UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Bayesian approach to Ecological Monitoring data: Latent Growth Curve Models

Pedro Miguel da Conceição Oliveira

Mestrado em Bioestatística

Trabalho de Projeto orientado por:
Prof.ª Doutora Patrícia de Zea Bermudez
Prof. Doutor José Lino Costa

2021

"We share our planet with millions of different species. Biodiversity is our safety net."
- World Wildlife Foudation

Acknowledgements

First of all I would like to thank my supervisors, Professor Patrícia de Zea Bermudez and Professor José Lino da Costa for all the help they provided in the design of this project and for pointing me in the right direction at the times when my path became more troubled. I would also like to thank the two institutions where this work was developed, the Faculty of Sciences of the University of Lisbon for the knowledge it has passed on to me over the years through its professors and employees, and the Marine and Environmental Sciences Center where I grew up as a person and as a professional through contact with new people and ways of working. In particular, I would like to thank the environmental monitoring team that collected the samples and compiled the data presented here, especially Gilda Silva who helped me with everything she could regarding the data and ideas for the design of this project.

A personal thank to the two fisherman, Carlos Pereira ("Quinchas") and Marco Fernandes ("Jacaré"), that helped with the determination of depth and distances of sampling stations.

For all the emotional and mental support I would also like to leave a huge thank you to my closest friends who over all these years have guided me through the most troublesome moments, namely Gilda Silva, Joana Cruz, Joshua Heumüller, Andreia Tracana, Helena Frasão and Margarida Raposo. On a more personal note, I want to thank my parents for the support they have given me all my life and especially in these last years when I started my academic path. To my biggest supporter ever, my grandfather, I want to dedicate all the work that presents itself here and I hope that, wherever he is, he can see how much I have grown up and dedicated myself to reaching this personal goal.

This work was supported by the Foundation for Science and Technology (FCT - *Fundação para a Ciência e Tecnologia*) through MARE (UID/MAR/04292/2019) and COASTNET (PINFRA/22128/2016).

Resumo

Os efeitos da influência antropogénica no meio ambiente têm-se manifestado, cada vez mais, como uma das principais razões que contribui para a alteração e redução da biodiversidade local e global. Face ao constante aumento populacional, é expectável que, se não forem adotadas medidas de cariz urgente, muitas espécies fiquem em risco ou em última instância, tais processos nefastos poderão conduzir à sua extinção.

Com vista a reverter este processo, ecologistas de todo o Mundo têm vindo a trabalhar continuamente no sentido de desenvolver mecanismos para deteção de riscos variados. Apoios institucionais têm-se revelado fulcrais no decorrer da implementação destes planos, providenciados tanto por autoridades de mitigação, como por entidades de proteção ambiental. No entanto, é importante salientar o papel fulcral da intervenção das comunidades locais.

Uma das principais causas da destruição maciça de habitats a nível global tem por base a emissão de substâncias de caráter poluente no meio ambiente. Dentro dos diferentes tipos de substâncias prejudiciais, a matéria orgânica proveniente de descargas de efluentes não tratados em áreas urbanas assume um papel importante na perturbação do equilíbrio e da estabilidade dos ecossistemas circundantes. Por forma a reduzir este tipo de contaminação nos sistemas aquáticos, induzida pela influência antropogénica, mecanismos de tratamento de águas urbanas e industriais têm vindo a ser desenvolvidos e aperfeiçoados, bem como as infraestruturas responsáveis por estes mesmos tratamentos, conhecidas como Estações de Tratamento de Águas Residuais (ETAR).

De forma a proceder à monitorização da atividade destas estações especializadas, diferentes planos foram desenhados, implementados e melhorados para avaliar se ao longo do tempo se verifica um efeito prejudicial da construção e funcionamento das infraestruturas supracitadas no meio ambiente. Várias colaborações entre instituições governamentais e especialistas ambientais têm vindo a ser feitas de forma a dar resposta a este tipo de problemas. Mais concretamente, em 2001 deu-se inicio a uma parceria entre a Câmara Municipal de Almada e o Centro de Ciências do Mar e do Ambiente (MARE), com o principal objetivo de avaliar o impacto ambiental da construção de ETARs municipais drenantes no estuário do Tejo nas comunidades ribeirinhas do concelho de Almada.

Em 2003, foi construída uma ETAR no Portinho da Costa, com o propósito de desativar um emissário de efluentes não tratados localizado no Porto do Buxo. Tendo sido obtidos dados relativos a este sistema, o seu tratamento e análise revelou-se necessário. Contudo, os métodos utilizados até então apenas descreviam as principais conclusões possíveis de deduzir, bem como a determinação da qualidade ecológica das águas numa janela temporal restrita através da utilização de um índice de qualidade ecológica. Um índice usado com bastante frequência na Europa é o Índice Biótico Marinho (AMBI - *AZTI's Marine Biotic Index*), que se baseia em dois conceitos principais para caracterizar os locais de amostragem, Coeficiente Biótico e Índice Biótico.

Com a acumulação de dados ao longo dos anos, para os diferentes locais de amostragem definidos

previamente, tornou-se possível uma análise longitudinal dos dados, que permite a descrever tendências e a quantificar alterações nas comunidades biológicas existentes. Neste trabalho serão utilizados estes dados recolhidos entre 2004 e 2011 (completando 8 anos de dados) para a descrição de três importantes variáveis biológicas (Abundância Total de organismos, Riqueza Taxonómica e Coeficiente Biótico).

Os Modelos Latentes de Curvas de Crescimento são uma ferramenta estatística frequentemente utilizada na análise de dados longitudinais, uma vez que permitem descrever e quantificar as alterações de uma determinada variável ao longo do tempo. Apesar destes modelos serem maioritariamente usados em Estatística Clássica, têm vindo a ser desenvolvidas abordagens do ponto de vista bayesiano.

À luz desta forma de pensamento estatístico, os dados observados são considerados como entidades fixas e os parâmetros são variáveis aleatórias, que possuem a sua própria distribuição de probabilidade. Para a aplicação do método bayesiano, é necessário atribuir uma distribuição inicial aos parâmetros, distribuição *a priori*, através da informação que se tem até ao momento. A partir desta, e em conjugação com a função de verosimilhança dos dados, uma outra distribuição, denominada distribuição *a posteriori*, pode ser obtida por meio do Teorema de Bayes. Esta última, fornece a informação renovada acerca dos parâmetros do modelo, após se terem observado os dados. Habitualmente a distribuição *a posteriori* dos parâmetros é bastante complexa, tendo, por isso, que se recorrer a metodologias de simulação estocástica, tais como o método de Metropolis-Hastings e o método de amostragem de Gibbs, ambos baseados no conceito de cadeias de Markov.

Para a aplicação dos Modelos Latentes de Curvas de Crescimento, é necessário ter a medição repetida de uma variável de interesse ao longo do tempo para a mesma unidade experimental. Com estas medições podem ser estimados o nível inicial (L) e o declive inicial (S) para cada unidade experimental. O caráter diferencial desta metodologia quando comparada com os Modelos Lineares Generalizados é a inclusão de um parâmetro extra denominado parâmetro de forma (α). A simplicidade destes modelos permite uma descrição detalhada da variável de interesse utilizando os parâmetros latentes (não observados) acima mencionados.

Para a realização deste trabalho foram selecionados os dados referentes a dois locais, Porto do Buxo e Portinho da Costa. Nestes locais, foram definidos radiais e transetos de forma a cobrir uma maior área. Como consequência desta decisão obtiveram-se 9 estações de amostragem no Porto do Buxo (3 transetos e 3 radiais) e 15 no Portinho da Costa (5 transetos e 3 radiais), perfazendo um total de 24 estações de amostragem.

Após uma análise detalhada dos dados, algumas decisões foram tomadas com o intuíto de obter resultados mais adequados. Resumidamente, de forma a eliminar o enviesamento provocado por um evento externo (alteração não explicada que ocorreu de igual forma para todas as estações de amostragem) identificado em 2007, for feito um corte aos dados iniciais ficando apenas com os últimos quatro anos (2008 a 2011). Além disso, foram separadas as estações do ano de forma a eliminar o efeito da sazonalidade presente nos dados, ou seja, para cada uma das variáveis de interesse, foram ajustados quatro modelos diferentes, um para cada estação do ano.

Para o ajustamento dos modelos, foram geradas 5 cadeias de Markov para cada parâmetro. De forma a não influenciar muito o resultado das simulações, foram utilizadas distribuições *a priori* vagas para que se pudesse deixar que "os dados falem". Relativamente aos resultados obtidos, de uma forma geral pode-se dizer que, ao longo dos quatro anos em estudo e para todas as estações do ano, os modelos propostos ajustaram-se bem em relação às tendências observadas. Resumidamente, para a abundância total de organismos e a riqueza taxonómica em cada estação do ano e em todas as estações de amostragem, é possível encontrar uma tendência crescente bem definida, com alguma variação em termos do declive.

Embora isto pudesse significar que as comunidades biológicas estariam a caminhar para uma melhor saúde no geral, tal não se verifica, uma vez que foi possível encontrar uma tendência crescente no coeficiente biótico, ainda que com um declive reduzido.

Por fim, com este trabalho foi possível avaliar a capacidade de ajustamento destes modelos a dados ecológicos. De um modo geral, estes modelos são úteis para a descrição e previsão de tendências em variáveis biológicas. No entanto, é necessário ter em atenção que, para que os resultados sejam o mais corretos possível, efeitos externos ao objeto principal da investigação devem ser atenuados ou eliminados, como foi o caso da sazonalidade. A possibilidade de inclusão de uma componente de sazonalidade no modelo não foi contemplada, ficando essa questão em aberto.

Palavras-chave: Inferência Bayesiana, Análise de Trajetórias, Dados Longitudinais, Monitorização Ecológica

Abstract

Anthropogenic pressure in the environment has been identified as a great cause of environmental destruction and biodiversity loss worldwide. Pollution by organic matter is one of the many pressures affecting the aquatic communities. With the objective of controlling the effects over these communities, Wastewater Treatment Plants have been developed and constructed near urban areas to reduce the amount of organic matter released to pristine ecosystems. In order to evaluate their functioning, ecological monitoring programs have been designed and implemented over the years. As a result, repeated measurements over the same experimental unit were obtained (longitudinal data). In 2003, the ecological monitoring of 24 sampling stations in Almada performed by the Marine and Environmental Science Center was initiated. This project has the main objective of analysing the data collected from 2004 to 2011 in this monitoring program using Latent Growth Curve Models (LGCM) to describe and quantify the changes in three biological variables (Total Abundance of organisms, Taxonomic Richness and Biotic Coefficient).

Keywords: Bayesian Inference, Latent Growth Curve Models, Longitudinal Data, Ecological Monitoring

Contents

Acknowledgements **v**

Resumo **vi**

Abstract **ix**

List of Tables **xiii**

List of Figures **xv**

1 Introduction **1**

1.1 Monitoring Aquatic Communities 2

1.1.1 Legislative Framework 2

1.1.2 Community Assessment 2

1.2 Wastewater Treatment Plants 3

1.3 Objectives and Outline 3

2 Data Background **5**

2.1 Study Area 5

2.2 Sampling Notes 6

2.2.1 Grain Size Analysis and Total Organic Matter (TOM) of Sediments 7

2.2.2 Abundance and Taxonomic richness 7

2.2.3 AZTI's Marine Biotic Index (AMBI) 7

2.3 Description of the Variables in Study 9

2.4 Exploratory Analysis 10

2.4.1 Overall Analysis 11

2.4.2 Annual Analysis 17

2.4.3 Season Analysis 18

2.4.4 Complete data set 20

3 Bayesian Methodology and Latent Growth Curve Models **22**

3.1 Bayes' Theorem 23

3.2 Prior distributions 25

3.3 Bayesian inference 26

3.3.1 Inference on posterior distributions . . . 26
3.3.2 Predictive distribution . . . 26
3.4 Markov Chains and the MCMC process . . . 27
3.5 Longitudinal data and the Latent Growth Curve Models . . . 27
3.5.1 Latent Growth Curve Model . . . 28
3.5.2 Inter-group variability analysis . . . 29

4 Application of the LGCMs to Ecological Monitoring data 31
4.1 The models . . . 31
4.2 Data description . . . 32
4.3 Trajectory characterization . . . 33
4.3.1 Convergence Assessment . . . 34
4.3.2 Variable mean description . . . 35
4.4 Inter-individual differences . . . 38
4.5 Sensitivity Analysis . . . 39

5 Final Remarks 40
5.1 Discussion . . . 40
5.2 Conclusion . . . 41
5.3 Future Work . . . 41

Bibliography 43

A Annual and Seasonal Influence tests' results 48

B Data Description 52
B.1 *Porto do Buxo* . . . 52
B.2 *Portinho da Costa* . . . 53

C LGCM's implemented in JAGS 54
C.1 *Total Abundance* and *Taxonomic Richness* . . . 54
C.2 *Biotic Coefficient* . . . 55

D Convergence Analysis 56
D.1 *Total Abundance* . . . 56
D.2 *Taxonomic Richness* . . . 62
D.3 *Biotic Coefficient* . . . 68

Abbreviations

ACF . . . . . Autocorrelation Function

AMBI . . . . . AZTI's Marine Biotic Index

BC . . . . . Biotic Coefficient

BI . . . . . Biotic Index

EG . . . . . Ecological Group

EU . . . . . European Union

ESS . . . . . Effective Sample Size

ETP . . . . . Equal Tail Probability

GDP . . . . . Gross Domestic Product

GPS . . . . . Global Positioning System

HPD . . . . . Highest Posterior Density

LGCM . . . . . Latent Growth Curve Model

MCMC . . . . . Markov Chain Monte Carlo

MGS . . . . . Mean Grain Size

PM . . . . . Posterior Mean

PSD . . . . . Posterior Standard Deviation

PSU . . . . . Practical Salinity Units

TOM . . . . . Total Organic Matter

WFD . . . . . Water Framework Directive

WWTP . . . . . Wastewater Treatment Plant

List of Tables

2.1 Sampling station dependent variables 6
2.2 Ecological groups and its designation 8
2.3 Correspondence between Biotic Index and Biotic Coefficient 9
2.4 Link functions and Mean Values 11
2.5 Wald's test results for transect and radial comparison of the biological variables in *Porto do Buxo* 12
2.6 Wald's test results for transect and radial comparison of the biological variables in *Portinho da Costa* 14
2.7 Wald's test results for transect and radial comparison of substrate variables 15
2.8 Wald's test results for places comparison of biological and substrate variables 16

4.1 Non-informative prior distributions 32
4.2 Posterior statistics of the parameters of the model for *Total Abundance* 36
4.3 Posterior statistics of the parameters of the model for *Taxonomic Richness* 37
4.4 Posterior statistics of the parameters of the model for *Biotic Coefficient* 37

A.1 Wald's test results for year comparison of biological variables 49
A.2 Wald's test results for year comparison of substrate variables 50
A.3 Wald's test results for season comparison of biological variables 50
A.4 Wald's test results for season comparison of substrate variables 51

B.1 Descriptive statistics of biological variables in *Porto do Buxo* 52
B.2 Descriptive statistics of biological variables in *Portinho da Costa* 53

D.1 Posterior statistics of the parameters of the model for *Total Abundance* 56
D.2 Posterior statistics of the parameters of the model for *Total Abundance* 56
D.3 Posterior statistics of the parameters of the model for *Total Abundance* 57
D.4 Posterior statistics of the parameters of the model for *Total Abundance* 57
D.5 Posterior statistics of the parameters of the model for *Taxonomic Richness* 62
D.6 Posterior statistics of the parameters of the model for *Taxonomic Richness* 62
D.7 Posterior statistics of the parameters of the model for *Taxonomic Richness* 62
D.8 Posterior statistics of the parameters of the model for *Taxonomic Richness* 63
D.9 Posterior statistics of the parameters of the model for *Biotic Coefficient* 68
D.10 Posterior statistics of the parameters of the model for *Biotic Coefficient* 68

D.11 Posterior statistics of the parameters of the model for *Biotic Coefficient* 68
D.12 Posterior statistics of the parameters of the model for *Biotic Coefficient* 69

List of Figures

2.1 Map of sampling location and design 7
2.2 Theoretical model for the distribution of the proportions of ecological groups, modified from Borja *et al.* (2000) 8
2.3 Box-plots for biological variables recorded in *Porto do Buxo* 12
2.4 Box-plots for biological variables recorded in *Porto do Buxo* 13
2.5 Box-plots for substrate variables recorded in *Porto do Buxo* and *Portinho da Costa* . . . 14
2.6 Box-plots for the biological variables recorded in *Porto do Buxo* and *Portinho da Costa* . 16
2.7 Box-plots of biological variables for year analysis 17
2.8 Box-plots of substrate variables for year analysis 18
2.9 Box-plots of biological variables for seasonality analysis 19
2.10 Box-plots of substrate variables for seasonality analysis 19
2.11 Plot of the complete data set 21

3.1 Growth curve model path diagram 30

4.1 Trajectory plot of the biological variables 33
4.2 Example of graphical analysis of chain convergence 34
4.3 Estimated and observed means 36
4.4 Person-specific estimates 39

D.1 Graphical representation of the convergence analysis 58
D.2 Graphical representation of the convergence analysis 59
D.3 Graphical representation of the convergence analysis 60
D.4 Graphical representation of the convergence analysis 61
D.5 Graphical representation of the convergence analysis 64
D.6 Graphical representation of the convergence analysis 65
D.7 Graphical representation of the convergence analysis 66
D.8 Graphical representation of the convergence analysis 67
D.9 Graphical representation of the convergence analysis 70
D.10 Graphical representation of the convergence analysis 71
D.11 Graphical representation of the convergence analysis 72
D.12 Graphical representation of the convergence analysis 73

1 Introduction

Human activities and population growth have been responsible for environmental degradation, decrease of ecological quality and biodiversity reduction [1, 2, 3]. Environment quality and stability were largely ignored by the decision-makers until the 1990s [4]. The relationship between economic growth and environmental quality had always been controversial. While some consider that an exclusive economical development leads to more environment degradation by ignoring environmental problems in exchange for profits, others assume that a consequence of the merely economic development is the contribute to environmental quality control [5] leading to the prevention of further ecosystem destruction. In fact, during the past decades, the gross domestic product (GDP) has increased greatly in most countries, but CO_2 emissions have also increased, indicating a continuous degradation following the economic growth trend [6].

Environmental conservation awareness began many years ago with the publication of 'Silent Spring' by Rachel Carson, in 1962 [7]. This book is a call for environmental protection and pollution reduction, presenting a different view of these problems to the reader. Following this, in 1997, a group of scientists published the "World Scientists' Warning to Humanity" where many subjects about environmental destruction were referred such as pollution in different habitats such as air, rivers, lakes and oceans, loss of soil productivity and others [8]. Although the efforts to warn populations about these problems, no major changes were implemented and the situation got worse. With natural resources reduction and the current problem of climate change, the European Commission developed the Europe 2020 program. In 2010 this document was written to identify new problems and also the actions that needed to be taken so that by 2020 the environmental quality could be improved and the natural resources problem would be at least minimized [9].

By the decade of 1970s, in the United States of America, some improvements were made in terms of engineering, by developing pollution control technologies [10]. Although it was assumed that these changes would contribute to a better quality of the environment, it was soon realized that it was not so. Trying to minimize further destruction, and ultimately stop and reverse it, a biomonitoring protocol was created [11] so that the conditions could be evaluated at any given time. To turn this task easier, ecologists had to find organisms that could help with their aim (bioindicator organisms). Macro invertebrates are widely used in these type of studies due to their responses to environmental changes [12, 13].

Macro invertebrates are animals characterized by the lack of a backbone that can be observed without the help of a microscope (considered to be the ones retained on a 500 μm screen) [14]. In ecology, these organisms are widely used due to their vast range of sensibilities to environmental disturbances, reduced or lack of mobility (many of them are sessile), short life cycles and, mainly, due to their substrate dependency, where there is a large accumulation of pollutants and the anthropogenic activities have a strong impact [13]. The combination of all these characteristics make benthic macro invertebrate a useful tool to evaluate human impacts in aquatic environments [15]. These communities tend to create patches where there

are usually good conditions for colonization. For this reason, the sampling of such communities should have this characteristic distribution into consideration, so that a great variety of patches are covered [16]. Patchiness can derive from environmental factors such as sediment composition, hydrodynamic conditions and not less relevant anthropogenic influence, such as organic enrichment of the environment.

1.1 Monitoring Aquatic Communities

1.1.1 Legislative Framework

Monitoring aquatic environments is an important process to allow the conservation of biological communities and to maintain the impact of anthropogenic activities under control. Intending to obtain a good water quality status in all water bodies, the European Union (EU) implemented the Water Framework Directive (WFD) [17]. This document establishes many concepts and actions that the Member States of the EU must adopt and apply to evaluate the pollution rates and sources in all their water bodies (fresh-water systems, transitional and coastal waters) as well as trying to recover degraded ecosystems. In the beginning of its application, although the actions were established, it was difficult to know what methods to use, so each country applied the WFD in its way [18, 19, 20]. Nowadays, an inter-calibration procedure has been implemented to standardise the methods to be applied.

EU regulation on urban wastewater treatment was developed in 1991 by the Council of the European Communities. This directive contains the concepts related to the concerning areas, schedule for infrastructure implementations (such as wastewater collectors and type of treatment before releasing the effluents into the environment), criteria for the classification of sensitive and non-sensitive areas (this classification ensures that more susceptible to fast degradation areas have the release of wastewaters undergoing a more strict treatment instead of a normal treatment like non-sensitive areas) and also the regulation to be applied in case of the non-fulfilment of the above mentioned rules [21].

In 1997, Portugal defined its' rules for wastewater treatment and discharge to aquatic environments [22]. Decree-law 152/97 determines all the conditions for wastewater treatments, its discharge and the monitoring needed to access the ecological impacts of these discharges. Sensitive and non-sensitive areas were determined to fulfil the requirements proposed by the European Council. This classification can be found in the second annex of Decree-law 152/97 (eutrophic water bodies or susceptible to eutrophication, surface freshwater destined to be used as potable water and areas where the conventional treatment is not enough to achieve the directive's plan).

1.1.2 Community Assessment

The continuous study of benthic communities, requires the full knowledge of how to sample them and what kind of methods should be applied to have a methodological procedure fixed throughout the years (to allow the comparison of different data sets). The major question is if a seasonal coverage is needed or not. Seasonal trends in these communities are a debatable topic since it depends not only on the geographical region but also on habitat's nature. Depending on the objective of the sampling process, seasonal coverage can be useful. Alden *et al.* (1997) [23] found that differences in terms of power for trend detection are not significant when a four-season, two-season or a single season sampling is considered. Other authors needed to have a four-season sampling method to be able to apply biological indices to evaluate ecological status through time [24, 25].

In terms of biological variables, two main values can be collected in each sample, the organisms'

abundance and the taxonomic richness. With these values, others can be calculated such as the ecological quality value of the sampled sites using indices like the AZTI's Marine Biotic Index (AMBI) proposed by Borja *et al.* (2000) [26]. This index informs about the pollution and health status of the sampled sites so that researchers can inform managers.

Since these organisms are substrate-dependent, knowledge about its composition is required because it can largely explain the presence or absence of some taxa. For that, a granulometric analysis is usually conducted to complement the biological part [16]. Organic enrichment is one of the most important pollution sources in these areas, so it is crucial to determinate the total organic matter present in the sediment as a way to quantify the approximate level of organic contamination [27]. Through the analysis of these monitoring values, the sampled sites can be characterized in terms of community changes over time and try to understand why did they occurred [28].

1.2 Wastewater Treatment Plants

The main purpose of Wastewater Treatment Plants (WWTP) is to reduce the concentration of contaminants in the effluents, so that the discharge of such waters does not have such a negative impact or to allow a reduced impact on the pristine ecosystem [29]. Although these processes are quite efficient for most contaminants, some solids can still be dissolved (biosolids) and, therefore, released to the natural environment. This matter, in high concentrations can become a contaminant for communities that live nearby the effluent [30]. Different methods have been developed to treat wastewaters and contemplates, at least two types of treatments, primary and secondary. These treatments involve physical-chemical and biological procedures, both processes including mud decantation [22]. There is still a third treatment that focuses on the removal of other contaminants and pathogenic agents.

There are three main purposes behind monitoring programs performed in this areas, general public health (make sure there are no harmful contaminants to humankind), environmental conservation (protect marine resources and ecosystems) and information for decision making by managers [31].

1.3 Objectives and Outline

In this work, data from a WWTP monitoring program in Almada will be used. This program started in 2001 with the assessment of the conditions before the construction of the WWTP in 2003. The sampling process and data gathering was performed by a specialized MARE's (Marine and Environmental Science Center) team. Data referent to the years of 2004 to 2011 will be used to study changes in biological variables such as the total abundance of organisms and taxonomic richness. Besides these biological parameters, the ecological water quality will also be studied using the AMBI tool as a quantification method. As mentioned above, analysis of the sediment composition is needed to better understand the characteristics and conservation status of the communities. For that same reason, trends in mean grain size and organic matter content will be described.

The main goal of this project is to implement Bayesian growth curve models (GCM) to understand and quantify the changes in the main variables of interest. Besides this methodological aim, the work has also the objective of comparing the communities present in two places affected by different anthropogenic pressures. The first place was affected by an untreated sewage outlet that was deactivated, so that the community should become healthier. In the other sampled place, the effects of the construction and functioning of a WWTP must be evaluated. Here, a slight decrease of organisms and ecological quality is

expected in the first moments, followed by a stabilization period. In more formal terms, the questions that are present to be answered by this work are the following:

- Are benthic communities different in the two places in terms of total abundance of organisms, taxonomic richness and water ecological quality?
- Does the WWTP have a negative effect on the biological communities around it?
- What are the main trends in the variables under study? And how do these trends tend to relate to the characteristics of the communities studied?
- Finally, are the proposed methods useful to analyse ecological monitoring data in general, or only for particular cases?

To proceed with the determined objectives, a first approach to the data is needed. For this reason, an introduction to the project, methods and data used is done in the following chapter, where a simple data visualization and statistical testing of trends was done using box-plots and Generalized Linear Models (Chapter 2). In chapter 3 the basics of Bayesian analyses and inference are shortly described, as well as the Latent Growth Curve Models' methodology that was used. The results of the above cited models can be found in chapter 4. Finally, the results discussion and final conclusions reached from a statistical and ecological point of view are presented in chapter 5, as well as some ideas for future work in this scientific area.

2 | Data Background

A partnership between the City council of Almada and the Centre of Oceanography (nowadays known as MARE) started in 2001 with the main purpose of characterizing and monitor the marine communities present in Almada's coastal area [32]. The project is divided in three distinct parts: (1) monitoring the benthic communities in *Portinho da Costa*, *Porto do Buxo* and *Mutela*; (2) determine heavy metal contaminations on the estuarine waters of Almada; and (3) characterization of the Beach Seine fishing activities on Almada's coastal zone.

Considering the first part of the project, resulting data from the monitoring program of *Portinho da Costa*'s WWTP and of the deactivated sewage outlet of *Porto do Buxo* since 2004 to 2011 will be used for the development of this work. Data collected from *Mutela*'s WWTP will not be used due to differences in geographical location, hydrodynamic conditions, and biological characteristics. Also, the obtained data from 2001 to 2002 will not be used due to major differences in sampling methods. These differences are the result of a change in the outlet's position from the projected site to the actual place. The data set is composed of 32 different observations of 24 pre-defined sampling stations (9 located in *Porto do Buxo* and 15 located in *Portinho da Costa*). The obtained observations are the result of a continuous monitoring process in these sites for an eight year period (2004-2011) and the sampling process occurred every trimester/season (see section 2.2 for more detail).

2.1 Study Area

Tagus Estuary is located on the West Coast of Portugal and it is one of the largest estuaries in Europe, having an area of 320 km^2 [33, 34]. It has a complex morphology, being narrow and deep near the mouth and transitioning to a broader and shallower system in an intermediate part, where tidal flats form, and finally narrowing in the upper part. In Almada the climate is temperate, characterized by dry summers, with temperatures ranging from 14 °C / 15 °C (Winter) to 26 °C / 28 °C (Summer) (relative to 2004-2011 period). This estuary has a salinity gradient that increases from upstream (30 PSU on Winter to 20 PSU on Summer) to downstream (34 PSU on Winter to 36 PSU on Summer) . The observed salinity range, from upstream to downstream could be explained by fresh-water inputs coming from Tagus River [35].

Sampling stations are situated near the south bank (Almada) of the Tagus Estuary (see Figure 2.1), at approximately 7 km from the river mouth. This place is characterized by a soft-bottom sediment mostly composed by sand (fine, medium and coarse). Proximity to river mouth determines the stability of salinity fluctuations throughout the year, except in some event such as the occurrence of floods.

2.2 Sampling Notes

The data set obtained with this monitoring program can be divided into three major time intervals: (1) pre-construction stage (December of 2001 until October of 2002), when the untreated sewage outlet in *Porto do Buxo* was directly emitting to the estuary, the WWTP was not yet built and the initial ecological status of biological communities was evaluated; (2) construction phase, where no samples were taken; and (3) operational phase (from March of 2004 until nowadays), when the sewage outlet of *Porto do Buxo* was deactivated and the WWTP was already full functioning. This continued sampling process made possible a temporal following of community evolution.

Samples were collected every trimester/season (March - Winter; June - Spring; September - Summer; December - Autumn) using a Day-like grab (modified Smith-McIntyre), with a 0.1 m^2 surface area. To evaluate the effect of proximity to outlets, a radial sampling design was preformed (stations in the same radial are at the same distance from the outlet; see Table 2.1). Different transects were chosen to cover more directions and a vaster area (see Figure 2.1). Posteriorly, a sorting process to separate organisms from sediment was performed and these were morphologically identified to the lowest taxonomic level possible, using dichotomous keys. Characterization of each species/taxa in terms of sensibility to organic enrichment was possible using a matrix provided by the AMBI software developed by the AZTI [36]. Simultaneously, sediment characteristics such as grain size analysis and total organic matter content were estimated.

Table 2.1: Observed values of sampling station dependent variables (Depth, Coastline Distance and Outlet Distance), defined by site, radial and transect.

Site	Radial	Transect	Depth (m)	Coastline Distance (m)	Outlet Distance (m)
Porto do Buxo	A	1	12	22.7	30
		2	12	40.0	30
		3	12	31.9	30
	B	1	10	38.2	90
		2	18	93.3	90
		3	18	51.5	90
	C	1	14	47.9	150
		2	20	153.2	150
		3	18	82.9	150
Portinho da Costa	A	1	28	182.5	20
		2	22	173.2	20
		3	18	146.1	20
		4	24	155.5	20
		5	24	173.3	20
	B	1	28	243.1	80
		2	28	195.6	80
		3	22	100.6	80
		4	26	97.8	80
		5	28	182.6	80
	C	1	30	296.2	140
		2	30	204.9	140
		3	24	94.1	140
		4	26	70.7	140
		5	30	189.1	140

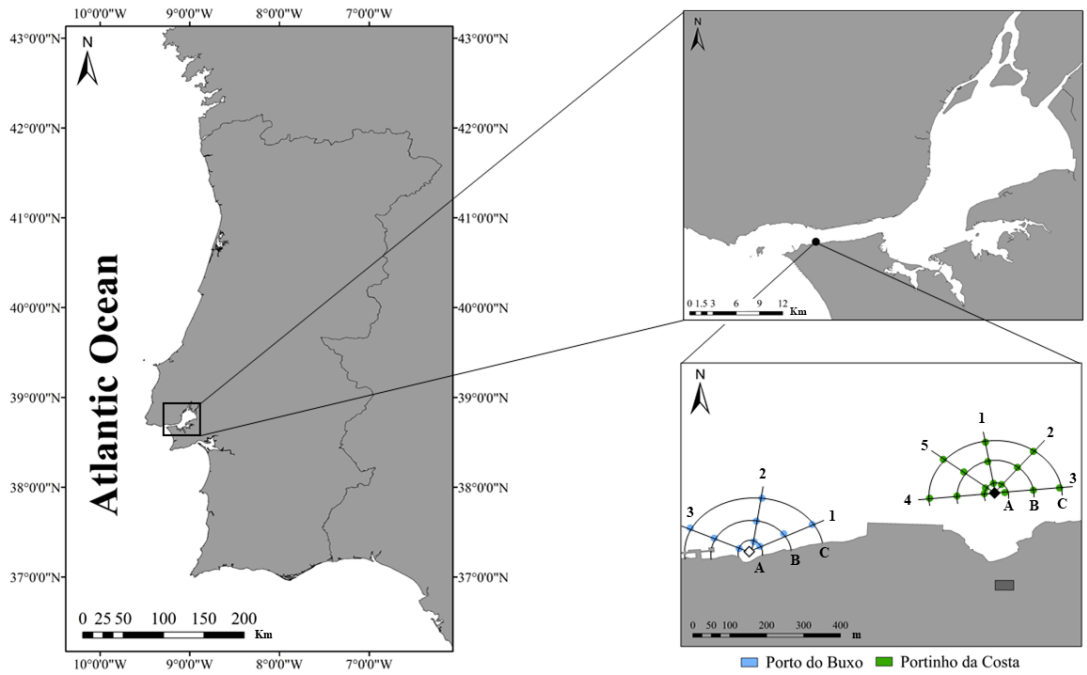

Figure 2.1: Map of sampling location and design , where ◇ represents a deactivated sewage outlet, ◆ the Wastewater Treatment Plant effluent, ● the sampling stations (blue ones from *Porto do Buxo* and the green ones from *Portinho da Costa*), curved lines the radials (points situated at the same distance from the outlet) and straight lines the transects (points situated at a same orientation) and ■ the Wastewater Treatment Plant's location.

2.2.1 Grain Size Analysis and Total Organic Matter (TOM) of Sediments

For grain analysis, approximately 100 g of sediment of each sample was dried at 60 °C and washed on a sieve with a 63 µm mesh, so that the finest fraction (mud) can be removed. The samples were posteriorly re-dried and passed through a French AFNOR sieve composed by four sieves with different mesh sizes (2 mm, 1.25 mm, 500 µm and 63 µm). The sample retained on each sieve was then weighted and the mud content obtained by subtraction.

To determine organic matter content in the samples, approximately 5 g of sediment were dried at 60 °C and combusted in a high-temperature furnace at 550 °C for a period of four hours. The proportion of organic matter in the sediments was calculated dividing the weight lost in combustion (difference between dry weight, W_d and the ash free dry weight, W_{af}) by the dry weight (Equation (2.1)).

$$TOM = \frac{W_d - W_{af}}{W_d} \tag{2.1}$$

2.2.2 Abundance and Taxonomic richness

The total abundance for one sampling stations was determined through the summation of all the individual abundances registered in the sorting and identification process. The same method was applied to the taxonomic richness, being the sum of the different taxa identified in each sampling station.

2.2.3 AZTI's Marine Biotic Index (AMBI)

This index is based on the concept of ecological groups (EG). Ecological groups are composed of taxa that share sensitivity levels to organic enrichment of the substrate they live on (a small description

of these ecological groups can be found in Table 2.2). This classification has been made and arranged throughout the years and a final updated list of taxa and their correspondent ecological group can be found in AZTI's web site together with the AMBI software [36].

Considering the theoretical model for the proportions' distribution of ecological groups on bays and beaches proposed by Hily *et al.* (1986) and latterly modified by Majeed (1987) [38, 39], Borja *et al.* (2000) [26] proposed a new generic theoretical model of these groups proportion's distribution (see Figure 2.2).

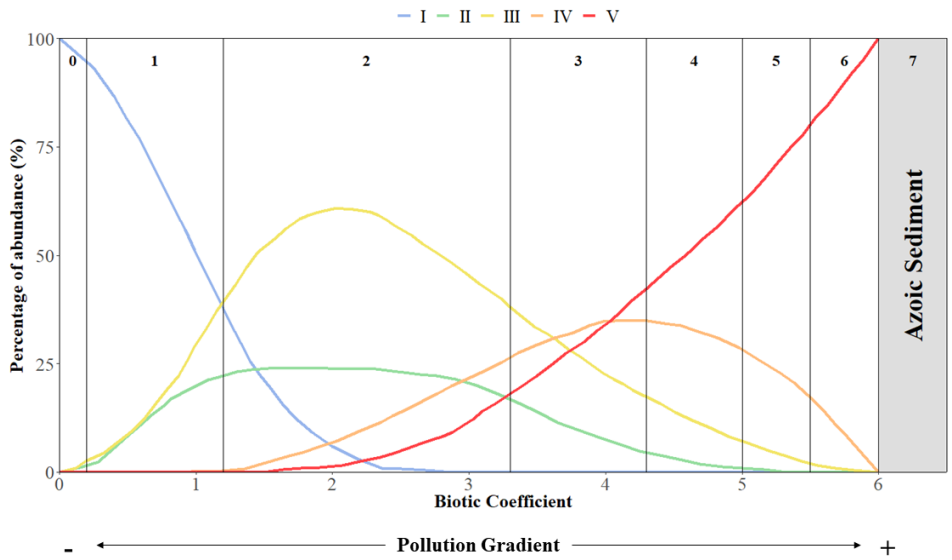

Figure 2.2: Theoretical model for the distribution of the proportions of ecological groups, modified from Borja *et al.* (2000) [26]. Separations marked by vertical lines represent the cut-off points for the discrete Biotic Index (indicated on the top part of the graph) of Hily *et al.* (1986), Majeed (1987) [38, 39] and their correspondence with the Biotic Coefficient values.

According to this model, a new index was developed to facilitate the decision making based in the determination of community ecological status and health. This new index is called AMBI and it is divided in two main parts. First, the calculation of the Biotic Coefficient (BC) using the proportions of abundance of each ecological group following Equation (2.2).

$$BC = (p_I \times 0) + (p_{II} \times 1.5) + (p_{III} \times 3) + (p_{IV} \times 4.5) + (p_V \times 6) \tag{2.2}$$

where p_I, p_{II}, p_{III}, p_{IV} and p_V represent the proportions for each EG. These proportions can be estimated

Table 2.2: Ecological groups designation, summarised by Grall and Glèmarec (1997) [37]

EG	Designation
I	Very sensitive to organic enrichment and present in normal conditions
II	Indifferent to organic enrichment, always present in low densities
III	Tolerant to excess organic enrichment, may occur always but these populations are stimulated by organic enrichment
IV	Second-order opportunistic species. Small ones with short life-cycles and adapted to live on reduced sediment
V	First-order opportunistic species. Mainly deposit feeders that proliferate in sediments reduced up to the surface
O	Non-evaluated species

by dividing the abundance of a given ecological group, n_{EG}, by the total abundance registered for that sample, n (Equation (2.3)).

$$\hat{p}_{EG} = \frac{n_{EG}}{n} \tag{2.3}$$

Secondly, the categorization of BC (real number between 0 and 6) is also relevant to assign a pollution classification and community's health to the sampling station. This attribution is based on the Biotic Index, a discrete categorical index that links the pollution status to the community's health. Borja *et al.* (2000) established a correspondence between these two indices for an easier classification of sampled sites (see Table 2.3).

Table 2.3: Correspondence between Biotic Index and Biotic Coefficient, adapted from Borja *et al.* (2000) [26].

Pollution Classification	Biotic Coefficient	Biotic Index	Community Health
Unpolluted	$0.0 < BC \leq 0.2$	0	Normal
Unpolluted	$0.2 < BC \leq 1.2$	1	Impoverish
Slightly Polluted	$1.2 < BC \leq 3.3$	2	Unbalanced
Meanly Polluted	$3.3 < BC \leq 4.3$	3	Transitional to Pollution
Meanly Polluted	$4.3 < BC \leq 5.0$	4	Polluted
Heavily Polluted	$5.0 < BC \leq 5.5$	5	Transitional to Heavy Pollution
Heavily Polluted	$5.5 < BC \leq 6.0$	6	Heavily Polluted
Extremely Polluted	Azoic [1]	7	Azoic

2.3 Description of the Variables in Study

Variables used in this study can be divided into four different groups: (1) biological variables, which give information about the community composition in certain areas; (2) substrate variables, that describe the soft-bottom in which these communities live in, (3) sampling station-dependent variables, the ones that are only related to the sampling station and not to the biological communities and (4) spatial-temporal variables.

To describe these communities, three biological variables were chosen: *taxonomic richness*, *total abundance*, and the obtained value for ecological quality using AMBI for each sampling station at each time. The first two variables are discrete representing counts, therefore assuming values equal or larger than 0. The third varies continuously between 0 and 6, assuming the value 7 when no organisms are found in a sample, meaning that the sampling station represents an azoic location.

In terms of substrate variables two were considered: the *mean grain size* (MGS) and the *total organic matter* (TOM) on sediments. Mean grain size was calculated using the package G2Sd from R software [40], resulting in a continuous variable that represents the mean diameter, in µm, of the substrate composing particles. Therefore, it is expected that they assume any positive value larger than 0. Total organic matter was calculated using Equation (2.1) and the result is a value between 0 and 1.

Sampling station-dependent variables are the ones that indicate the site's position in the estuary, being the *coastline distance*, determined using the measuring tool in Google Maps [41]. Outlet distances (specified in Table 2.1 and determined by the sampling design) and depth were measured using a GPS device of a Trafaria's fisherman.

Spatial-temporal variables were considered to support the explanation of the biological ones such as sampling season and position. Sampling season is a categorical variable assuming one of four possible

[1] non-existence of life; assumed when no organisms are found in a sample

values ("W" - Winter, "Sp" - Spring, "Su" - Summer, "A" - Autumn) and sampling place, a dichotomous variable indicating the local where a sample was collected ("B" - *Porto do Buxo*, "C" - *Portinho da Costa*).

2.4 Exploratory Analysis

For the exploratory analysis, the main data matrix was re-arranged into three different ones depending on the type of analysis that would be performed. The overall dataset, combining all data collected throughout the years to evaluate the major differences between sampling stations and sites. The annual dataset used to make an initial analysis of the time influence. The same process was applied to a seasonal dataset, to investigate the existence of a pattern.

Although these data may violate the assumption of independence between observations due to a time related sampling method, since the experimental design is balanced, the Generalized Linear Models (GLM) will be used as a way to more deeply describe our data. For the above reasons the results should be considered with extreme caution. These models will be fitted to 32 observations from 24 different sites (corresponding to a total of 768 data points). GLM will be used as a complementary approach to graphical analysis. In each step, estimated values for coefficients and their significance will be presented and interpreted.

Given that the *total abundance* and *taxonomic richness* are count variables with a high variance relative to the mean, the models were created using the negative binomial distribution. For the biotic coefficient/AMBI and mean grain size the normal distribution was considered, since there are no reasons to doubt this assumption. Finally, for TOM, as it is a proportion, the natural distribution is the beta distribution [42].

Generalized Linear Models

These type of models are an extension of the commonly used linear models [43] as they also consider a variable of interest, Y (response variable) with independent observations and one or more covariates, $\mathbf{X}$. For that reason, the Linear model's general form can be written as:

$$Y = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.4}$$

where $\mathbf{Z}$ is a $(n \times (p+1))$ specification matrix composed by a first vector of 1's and p covariate vectors, $\boldsymbol{\beta}$ is a parameter vector of length $p+1$ and $\boldsymbol{\varepsilon}$ is a vector of independent random errors. It should be noted that n represents the number of experimental units.

GLM's are considered an extension of linear models in two ways: (1) The distribution of the response variable does not have to be the Normal distribution, being able to be any one of the distributions from the Exponential family; (2) linearity between a function of the mean value of Y and the covariates $\mathbf{X}$ is maintained, with the link function being any differentiable function. For the response variables considered in this study, link functions and mean values are presented in Table 2.4.

For the comprehension of GLM methodology, there are two concepts worth describing. The first one is the linear predictor, as it is the common ground between GLM and Linear Models. The linear predictor, η_i, for individual i, with a set of covariates $(X_1, X_2, \ldots, X_p)$ is:

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

where β_0 represents the value of η_i when no covariates are added, β_1, β_2, $\ldots$, β_p are the coefficients

Table 2.4: Some Generalized Linear Models for continuous and discrete response variables, considering the distribution, link function and mean value (μ_i).

Response Variable's Distribution	Link Function	Mean Value
Normal	Identity	$\mathbf{z}_i^T\beta$
Beta	Logit	$\frac{exp(\mathbf{z}_i\beta)}{1+exp(\mathbf{z}_i\beta)}$
Negative Binomial	Logarithm	$exp(\mathbf{z}_i\beta)$

associated with the covariates ($X_1, X_2, \ldots, X_p$). The next concept that needs some explanation is the link function. This function is the one responsible for the connection between the linear predictor and the mean value of Y_i, such that:

$$\eta_i = g(\mu_i)$$

where μ_i is the mean value of Y_i and $g(.)$ is a differentiable function.

This modelling methods, considering one explanatory variable at a time, were used to better understand the data. Therefore, the model selection step of the GLM procedure is not required. A nullity test is applied to every model coefficient using Wald's test. Hypothesis in test are:

$$H_0 : \beta_j = 0 \quad vs. \quad H_1 : \beta_j \neq 0$$

Considering the results of the test, a decision will be made considering the significance level of 5%.

Given that, in this work, some independent variables are categorical with several categories, the problem of multiple testing will arise. In order to reduce the bias created by multiple comparisons, many corrections can be applied. In this case, a Bonferroni-type adjustment will be adopted [44]. Considering the general significance level of 5%, this Bonferroni-type adjustment consists in dividing the significance level by the total number of comparisons that were made. Specifically for this work, for the fitted models, there are 2, 3, 4 or 7 comparisons, which will lead to new significance levels of 2.50%, 1.67%, 1.25% and 0.70%, respectively.

2.4.1 Overall Analysis

In *Porto do Buxo*, in terms of *total abundance* and *taxonomic richness*, there seems to be a slight decreasing trend while the coastline distance increases, which is evident in transect 2. This pattern is very clear in Figure 2.3 (A) and (B), where, for transect 2, the values tend to decrease from radials A to C. In transects 1 and 3, this decreasing trend is not observed, but the coastline distance of points in these transects does not change as much (see Table 2.1). This trend can be emphasized by observing Wald's test results of the performed using models that consider the *coastline distance* as explanatory variable ($z = -4.138$ with $p - value \ll 0.05$ for *total abundance* as the response variable and $z = -2.293$ with $p - value < 0.025$ for *taxonomic richness* as the response variable). Considering the Biotic Coefficient values (Figure 2.3 (C)), all transects showed similar values in radials A and B, which means that near the pollution source there is a clear tendency for higher values, related to a decreased ecological quality. In radial C, lower values were observed reflecting a better quality status in the community. Considering radial A as the reference category, biotic coefficient values tend to be lower in 0.405 points in radial C

(significantly different from 0 showed by a $t = -3.503$ with $p - value < 0.025$ from the Wald's test).

Table 2.5: Values obtained for the coefficient estimates (top value) of all parameters involved and the test statistics (bottom left value) and respective p-value (bottom right value in parentheses) of Wald's test for the two models created for each one of the response variables (N - Total Abundance; S - Taxonomic Richness; BC - Biotic Coefficient). A total number of 288 observations were used in each model. As a reminder, given that there are 2 comparisons being made, the significance level is corrected to 2.5%.

		N	S	BC
(Intercept)		6.432 63.449 ($\ll 0.05$)	3.859 64.365 ($\ll 0.05$)	2.541 30.458 ($\ll 0.05$)
Transect	2	-0.519 -3.616 (< 0.025)	-0.136 -1.601 (> 0.025)	-0.108 -0.914 (> 0.025)
	3	0.030 0.213 (> 0.025)	0.062 0.733 (> 0.025)	0.033 0.279 (> 0.025)
(Intercept)		6.336 61.197 ($\ll 0.05$)	3.822 63.494 ($\ll 0.05$)	2.693 32.930 ($\ll 0.05$)
Radial	B	0.042 0.288 (> 0.025)	0.106 1.256 (> 0.025)	-0.125 -1.080 (> 0.025)
	C	-0.160 -1.092 (> 0.025)	-0.063 -0.746 (> 0.025)	-0.405 -3.503 (< 0.025)

To understand the behaviour of these variables in the different transects and radials, models were built considering these as explanatory variables, one at each time. Wald's test results can be found in Table 2.5 where, for each variable, the estimated coefficient value, the observed test's statistics and the associated *p-value* are presented. Resuming this information, transect 2 is the one that is separated from the other two in terms of individuals' abundance, although it is not differentiated in terms of taxonomic richness. As for community's health, there are no significant differences between transects 2 and 3 and the

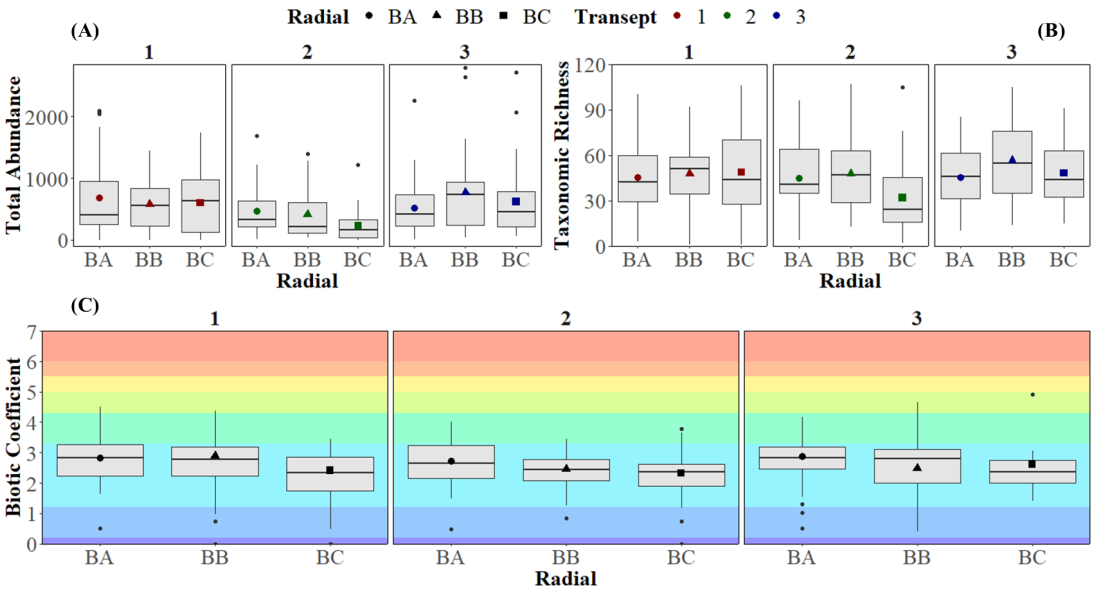

Figure 2.3: Box-plots for biological variables recorded in *Porto do Buxo* between 2004 and 2011. (A) Total abundance; (B) Taxonomic richness; and (C) Biotic coefficient calculated using Equation (2.1), where different colours separate each value of the Biotic Index described in Table 2.3. Dots with different shapes represent the observed mean for that sampling station. Each graph is separated in three, for each transect (represented on the top part of the graph) and in each transect, values for each radial are presented.

reference category, but, radial C tends to significantly lower values, when compared to radial A (showed by the negative sign of the estimated coefficient value).

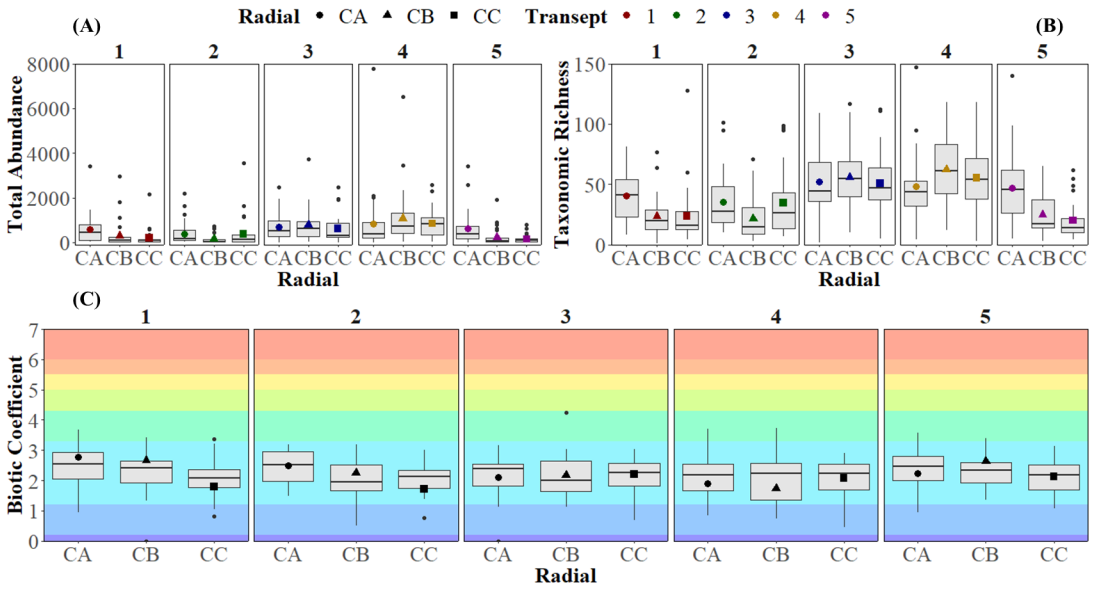

Figure 2.4: Box-plots for biological variables recorded in *Portinho da Costa* between 2004 and 2011. (A) Total abundance; (B) Taxonomic richness; and (C) Biotic coefficient calculated using Equation (2.1), where different colours separate each value of the Biotic Index described in Table 2.3. Dots with different shapes represent the observed mean for that sampling station. Each graph is separated in three, for each transect (represented on the top part of the graph) and in each transect, values for each radial are presented.

In *Portinho da Costa*, the same trend seems to appear, *i.e.*, transects that have all sites close to the coastline (3 and 4), are the ones with higher *total abundance* and *taxonomic richness* (see Figure 2.4 A and B). In the other three transects (1, 2 and 5), located further away from the coast (from radial A to C), the values taken by these variables tend to decrease. These tendencies are reflected in the results of Wald's test for the models fitted to each of the response variables and considering the *coastline distance* as the explanatory variable ($z = -8.043$ with a $p - value \ll 0.05$ for *total abundance*; $z = -9.324$ with a $p - value \ll 0.05$ for *taxonomic richness*). *BC* values in *Portinho da Costa*, as in *Porto do Buxo*, do not have a direct relationship with the *distance to the coastline* but it has a significant decline with the distance to the pollution source (decreasing estimated values from radial A to C, see Table 2.6).

As done for *Porto do Buxo*, the influence of transects and radials was tested using models created that consider these two variables as the independent ones. These results can be found in Table 2.6, where the test statistics and p-values are presented. Summarizing this information, it is worth noting that transects 3 and 4 are the most different in terms of abundance and number of taxa, having a tendency to have more organisms and taxonomic richness than the reference category. Unlike *Porto do Buxo*, radials seem to have an influence on all biological variables as there is a significant decrease of taxa and organisms from radial A to C, as well as a decrease in BC values.

In Figure 2.5, it is possible to observe the values of mean grain size and total organic matter. With regard to the mean grain size, the same trend mentioned above can be observed in both places. Along with the increasing distance from the coast, the grain size tends to be smaller. A small grain size represents a higher percentage of mud or fine sand in the substrate. This trend was again proven using generalized linear models, where *coastline distance* was used as an explanatory variable and mean grain size as a response ($t = -6.776$ with a $p - value \ll 0.01$ in *Porto do Buxo*; $t = -7.595$ with a $p - value \ll 0.01$ in

Table 2.6: Values obtained for the coefficient estimates (top value) of all parameters involved and the test statistics (bottom left value) and respective p-value (bottom right value in parentheses) of Wald's test for the two models created for each one of the response variables (N - Total Abundance; S - Taxonomic Richness; BC - Biotic Coefficient). A total number of 480 observations were used in each model. As a reminder, given that there are 4 and 2 comparisons being made, the significance level is corrected to 1.25% and 2.5%, respectively.

		N	S	BC
(Intercept)		5.897 47.932 ($\ll$ 0.05)	3.378 47.945 ($\ll$ 0.05)	2.267 33.953 ($\ll$ 0.05)
Transect	2	-0.183 -1.055 (> 0.0125)	0.048 0.481 (> 0.0125)	-0.090 -0.956 (> 0.0125)
	3	0.656 3.779 ($\ll$ 0.0125)	0.590 5.984 ($\ll$ 0.0125)	-0.132 -1.402 (> 0.0125)
	4	0.928 5.227 ($\ll$ 0.0125)	0.636 6.316 ($\ll$ 0.0125)	-0.162 -1.681 (> 0.0125)
	5	-0.088 -0.510 (> 0.0125)	0.041 0.408 (> 0.0125)	-0.006 -0.065 (> 0.0125)
(Intercept)		6.434 64.525 ($\ll$ 0.05)	3.794 65.369 ($\ll$ 0.05)	2.322 45.257 ($\ll$ 0.05)
Radial	B	-0.217 -1.537 (> 0.025)	-0.173 -2.098 (< 0.025)	-0.164 -2.263 (< 0.025)
	C	-0.377 -2.670 (< 0.025)	-0.191 -2.098 (< 0.025)	-0.233 -3.205 (< 0.025)

Portinho da Costa). In *Porto do Buxo*, there is no significant changes in total organic matter across all radials compared to radial A, although transect 3 tends to have approximately 0.38% (estimated value obtained using the fitted model) less organic matter in its sediments (see Table 2.7). In *Portinho da Costa*, values are higher in radial A, radial B shows a tendency to present less 0.53% organic matter and radial C less 0.38%. In transect 4, TOM values are significantly higher than in the other transects ($z = 2.965$ with a $p - value < 0.01$).

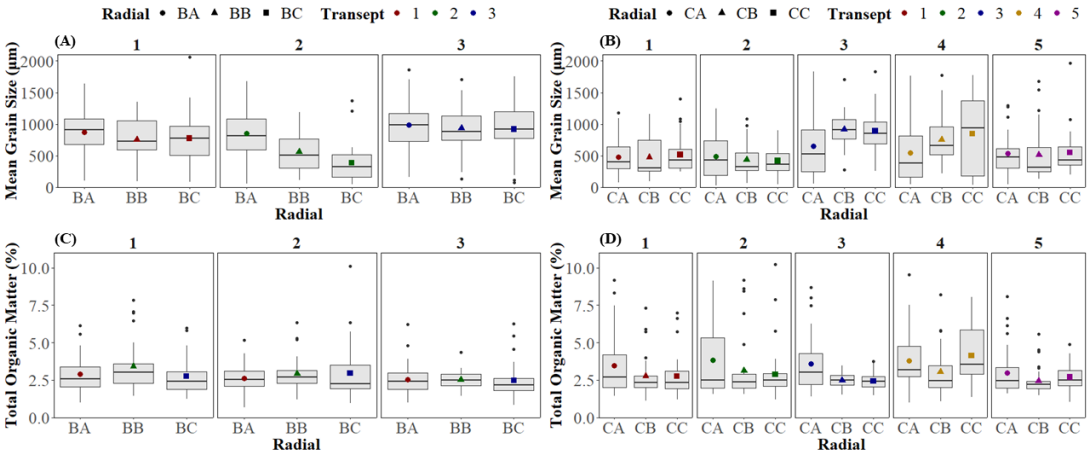

Figure 2.5: Box-plots for substrate variables recorded in *Porto do Buxo* and *Portinho da Costa* between 2004 and 2011. (A) And (B) Mean Grain Size; (C) and (D) Total Organic Matter. Dots with different shapes represent the mean observed for that sampling station. Each graph is separated in three or five, for each transect (represented on the top part of the graph) and in each transect the values for each radial are presented.

In terms of grain size, *Porto do Buxo* has a larger grain size than *Portinho da Costa*. Although both sites are characterised by having a sandy bottom, sand grains in *Porto do Buxo* are bigger than the ones from *Portinho da Costa*. It may seem an irrelevant characteristic, but it is possible to conclude that, the

smaller mean grain size of the sediment increases the chance of organic matter retention. This can be observed by a significant negative correlation between these two variables ($r = -0.227$, $p - value \ll 0.01$).

Table 2.7: Values obtained for coefficient estimates (top value) of all parameters involved and test statistics (bottom left value) and respective p-value (bottom right value in parentheses) of the Wald's test for the two models created for each one of the response variables (MGS - Mean Grain Size; TOM - Total Organic Matter). A total number of 288 and 480 observations were used for models in *Porto do Buxo* and *Portinho da Costa*, respectively. As a reminder, given that there are 2 and 4 comparisons being made, the significance level is corrected to 2.5% and 1.25%, respectively.

		Porto do Buxo		*Portinho da Costa*	
		MGS	TOM	MGS	TOM
(Intercept)		798.03 20.719 ($\ll$ 0.05)	-3.482 -87.672 ($\ll$ 0.05)	490.22 12.884 ($\ll$ 0.05)	-3.477 -74.224 ($\ll$ 0.05)
Transect	2	-198.42 -3.643 ($\ll$ 0.025)	-0.061 -1.093 (> 0.025)	-42.60 -0.792 (> 0.0125)	0.045 0.695 (> 0.0125)
	3	147.95 2.709 (< 0.025)	-0.141 -2.466 (< 0.025)	327.31 6.083 ($\ll$ 0.0125)	-0.013 -0.198 (> 0.0125)
	4			221.96 4.070 ($\ll$ 0.0125)	0.189 2.965 (< 0.0125)
	5			43.28 0.804 (> 0.0125)	-0.060 -0.906 (> 0.0125)
(Intercept)		903.08 22.376 ($\ll$ 0.05)	-3.577 -86.139 ($\ll$ 0.05)	537.71 17.139 ($\ll$ 0.05)	-3.346 -96.207 ($\ll$ 0.05)
Radial	B	-155.74 -2.736 (< 0.025)	0.098 1.729 (> 0.025)	80.48 1.811 (> 0.025)	-0.176 -3.488 ($\ll$ 0.025)
	C	-210.32 -3.695 ($\ll$ 0.025)	-0.016 -0.278 (> 0.025)	103.86 2.337 (< 0.025)	-0.123 -2.466 (< 0.025)

Comparison between the two places sampled is needed to fully understand the effect of the presence of a WWTP in *Portinho da Costa*. With all sampling stations studied and compared, what was left to do was compare the two places, since there is a main difference between them (the existence of an input of treated residual waters in *Portinho da Costa*). For this comparison, the very same methodology was applied, specifically the use of generalized linear models for all six variables studied above, but this time with the location as the explanatory variable and box-plots for visualization.

Looking at the abundance's box-plot, it can be seen that the means do not seem to differ (Figure 2.6 (A) blue and green dots), although in *Portinho da Costa* there is a larger dispersion than in *Porto do Buxo*. Wald's test results, considering the *total abundance* as the dependent variable and the categorical variable *local* as the independent one, supports the previous statement (all results can be found in Table 2.8). *Taxonomic richness* in *Portinho da Costa* seems to be lower on average (Figure 2.6 (B)) and this difference was proved to be significant through the results of Wald's test. It is worth to emphasize that it was in this place that higher values of *taxonomic richness* were recorded. Analysing the results obtained for AMBI (Figure 2.6 (C)), once again *Portinho da Costa* have shown significantly lower values (less 0.326 points on average), translating in a better environmental quality than in *Porto do Buxo*.

As mentioned before, *Porto do Buxo* tends to present larger grain size than *Portinho da Costa*, this trend has proved to be significant (see Table 2.8), as it can be seen by looking at Wald's tests results. When comparing both places as a whole, the negative relation between MGS and TOM (Figure 2.6 (D) and (E)) stands out since *Porto do Buxo* tends to have a larger grain size (approximately 181 μm on average) and less organic matter (approximately 0.2 % on average, value obtained using the fitted model) and the opposite is observed in *Portinho da Costa*.

Table 2.8: Values obtained for coefficient estimates (top value) of all parameters involved and test statistics (bottom left value) and respective p-value (bottom right value in parentheses) of the Wald's test for the models created for each one of the response variables (N - Total Abundance; S - Taxonomic Richness; BC - Biotic Coefficient; MGS - Mean Grain Size; TOM - Total Organic Matter; P - *Portinho da Costa*). A total number of 768 observations were used in each model.

		N	S	BC
(Intercept)		6.300 91.227 ($\ll$ 0.05)	3.839 95.524 ($\ll$ 0.05)	2.516 59.637 ($\ll$ 0.05)
Site	P	-0.052 -0.589 (> 0.025)	-0.162 -3.172 (< 0.025)	-0.326 -6.086 ($\ll$ 0.025)
		MGS		TOM
(Intercept)		780.64 33.139 ($\ll$ 0.05)		-3.525 -132.017 ($\ll$ 0.05)
Site	P	-181.60 -6.087 ($\ll$ 0.025)		0.070 2.119 (< 0.025)

After comparing all sampling stations and places between them considering the biological and substrate variables, the time influence needs to be studied. In sections 2.4.2 and 2.4.3 the analysis of annual and seasonal influence will be presented. This analysis was performed following the same methods mentioned above (using box-plots for data visualization and generalized linear models for testing the effect of the independent variable on the outcome variable) and for each place separately, since it was verified that there were significant differences between them. The complete tables with the estimated coefficients and Wald's test results are found in Appendix A.

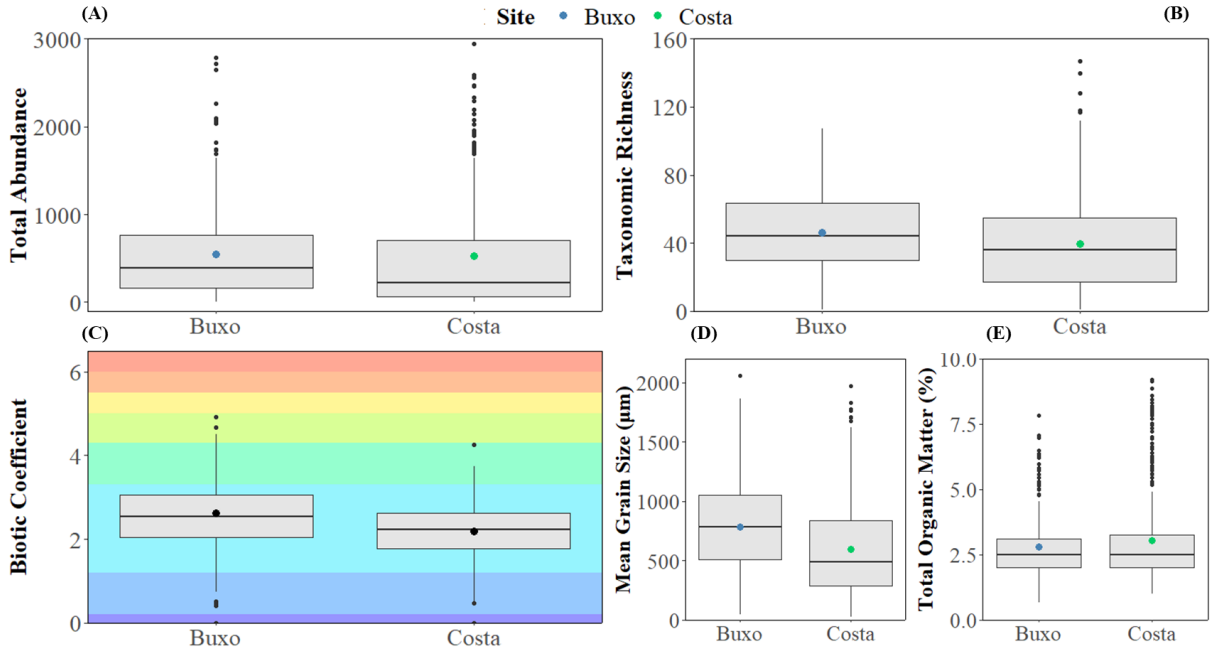

Figure 2.6: Box-plots for the biological variables recorded in *Porto do Buxo* and *Portinho da Costa* between 2004 and 2011. (A) Total abundance; (B) Taxonomic richness; (C) Biotic coefficient, where the different colours separate each value of the Biotic Index described in Table 2.3; (D) Mean Grain Size; and (F) Total Organic Matter. Dots with different colours represent the mean observed for that sampling station.

2.4.2 Annual Analysis

An increasing number of taxa and organisms starting in 2004 was found in both places until 2007, as well as a decreasing trend in BC values, meaning that an improvement on communities' health might have occurred (see Figure 2.7). In 2008, it is possible that some sort of unknown event happened in both communities and that might have induced the disappearing of many taxa, conducting to a repercussion on the ecological quality. Even though *taxonomic richness* and *total abundance* started to increase from 2008 until 2011, BC values have also increased, reflecting a worse health status. Therefore, generalized linear models were built for three different time periods (2004-2007, 2007-2008 and 2008-2011) considering the year variable as discrete categorical and the first year of each time interval as the reference category.

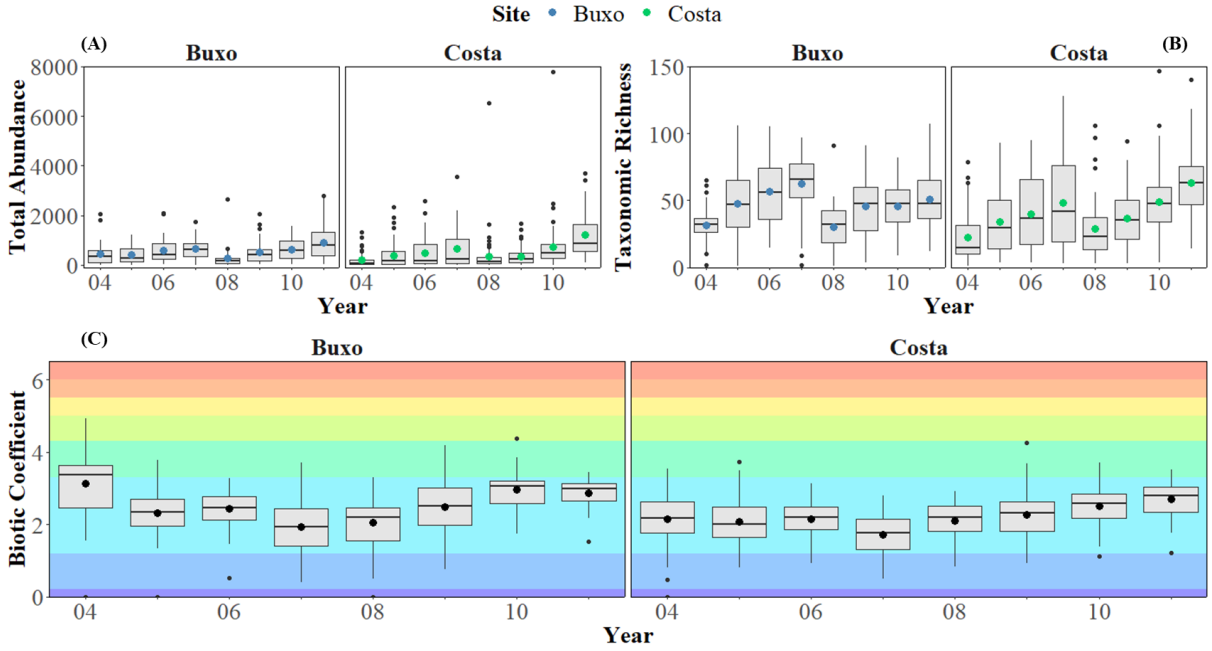

Figure 2.7: Box-plots of biological variables recorded in *Porto do Buxo* (PB) and *Portinho da Costa* (PC) between 2004 (04) and 2011 (11). (A) Total abundance; (B) Taxonomic richness; (C) Biotic coefficient, where the different colours separate each value of the Biotic Index described in Table 2.3. Blue and Green dots represent the mean observed for that place in the year considered. Each graph is separated in two, for each site (represented on the top part of the graph).

Considering the obtained results concerning *Porto do Buxo*, for *total abundance* in the first time interval, significant differences were not found (see Table A.1), although it should be noted the presence of a slight increasing trend. This trend is also accompanied by a significant increase of taxa. From 2007 to 2008, a significant drop was found, since the estimated value for *total abundance* was approximately 656 individuals (belonging to 63 different taxa) in 2007. In 2008 the estimate was approximately 261 individuals (belonging to 30 different taxa), reflecting on a 60 % loss of organisms and 55 % loss of taxa. In the third period considered for this work, a great recovery on these values can be seen and it is significantly higher every year. As described above, BC values tended to decrease in the first period and this descendent trend was proved to be significant, as well as the shown increasing trend in the third period.

Total Abundance in *Portinho da Costa*, had also significantly increased during first period (see Table A.1) and a significant loss of approximately 46 % of all organisms (from an average of 647 in 2007 to an average of 301 in 2008), during the second time interval (less than in *Porto do Buxo*). This loss was accompanied with a clear reduction of 59 % (from an average of 49 taxa in 2007 to an average of 20

taxa in 2008) on present taxa. For the third period, a great community recovery was observed, reaching almost the same number of taxa and organisms before the drop that occurred in 2008. Although the community recovered in numerical terms, taxa found there tended to be from higher ecological groups (more opportunistic taxa of first and second order), represented by an increasing trend of the *Biotic Coefficient*, observed from 2008 until 2011.

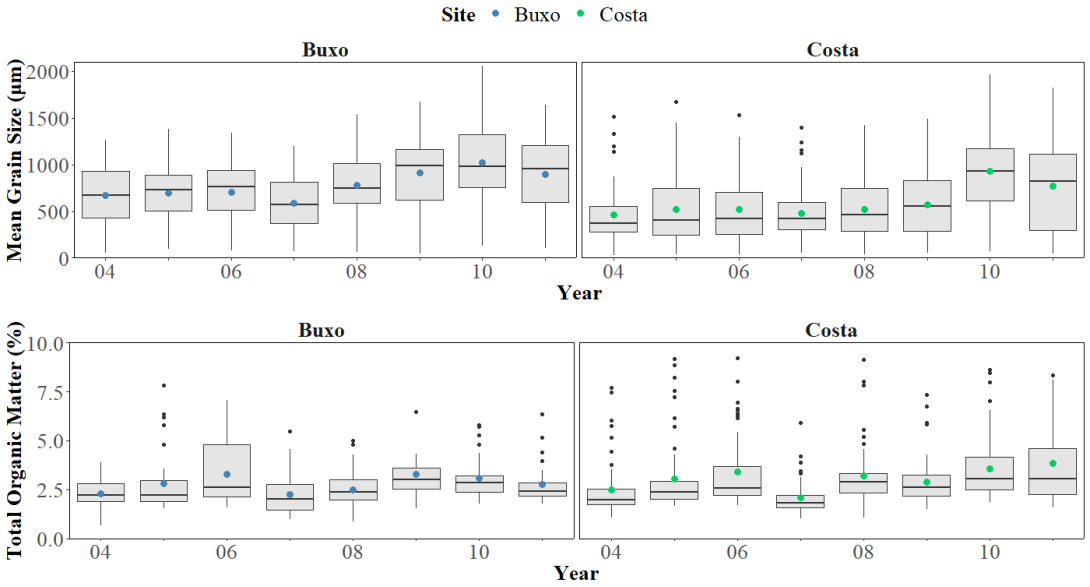

Figure 2.8: Box-plots of substrate variables recorded in *Porto do Buxo* (PB) and *Portinho da Costa* (PC) between 2004 (04) and 2011 (11). (A) Mean Grain Size; (B) Total Organic Matter. Blue and Green dots represent the mean observed for that place in the year considered. Each graph is separated in two, for each site (represented on the top part of the graph).

Since the observed trend in biological variables was not found in substrate variables (see Figure 2.8), the main analysis were performed without the time intervals considered before. Considering the substrate variables, major changes in *mean grain size* (compared to 2004) were not found until 2009 (see Table A.2), meaning that this variable does not seem to be related with the phenomenon that affected the biological variables. This fact was inferred considering that in the same years where there was a significant increase (2004 to 2007) and decline (2007 - 2008) in all variables related with the community's composition. In the last two years (2010-2011) there was a significantly higher presence of coarser sand in *Porto do Buxo* (larger than 1000 µm) and medium to coarse sand in *Portinho da Costa* (between 500 and 2000 µm). Considering *total organic matter*, 2006 and 2009 to 2011 were the years that showed a significantly higher value of organic matter in both sites (compared with the reference category). Although in *Porto do Buxo* from 2009 to 2011 the observed tendency is to return to values similar to the ones observed in 2004 and in *Portinho da Costa* the tendency is an increase of organic matter amount in sediments.

2.4.3 Season Analysis

In *Porto do Buxo*, *total abundance* and *taxonomic richness* share a rising trend from Winter to Autumn (see Figure 2.9 (A) and (B)). These increasing values were proven to be significant between Winter and Autumn, regarding *total abundance* and between Winter and all the other seasons regarding *taxonomic richness* (see Table A.3). As for the ecological quality measurement, higher values tend to appear in Winter and lower values tend to appear in Summer (see Figure 2.9 (C)). Differences between these two seasons were proven to be significant for both sampled places.

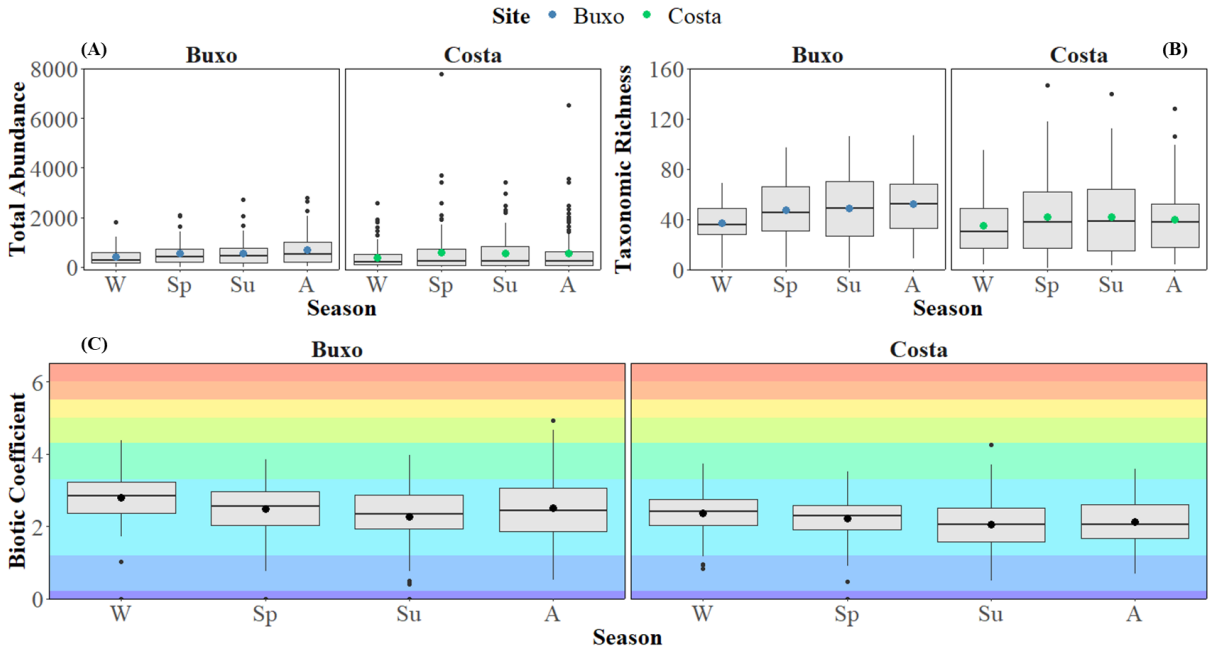

Figure 2.9: Box-plots of biological variables recorded in *Porto do Buxo* (PB) and *Portinho da Costa* (PC) between 2004 and 2011. (A) Total abundance; (B) Taxonomic richness; (C) Biotic coefficient, where the different colours separate each value of the Biotic Index described in Table 2.3. Blue and Green dots represent the mean observed for that place in the season considered. Each graph is separated in two, for each place (represented on the top part of the graph).

There were no significant differences found when analysing both substrate variables (see Table A.4). This can be observed in Figure 2.10, as average values are similar all year round.

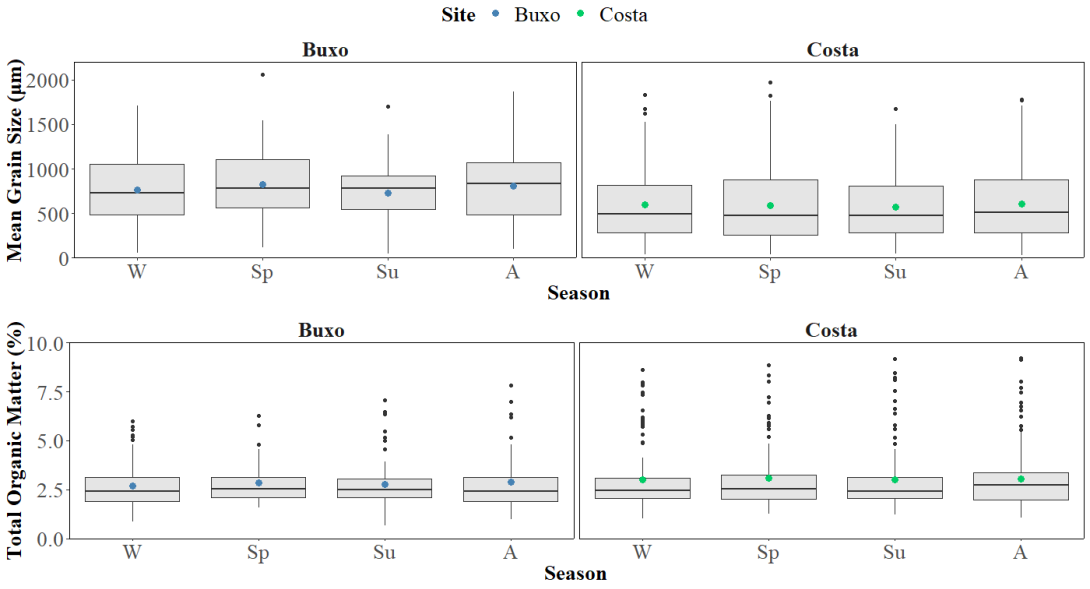

Figure 2.10: Box-plots of substrate variables recorded in *Porto do Buxo* (PB) and *Portinho da Costa* (PC) between 2004 and 2011. (A) Mean Grain Size; (B) Total Organic Matter. Blue and Green dots represent the mean observed for that place in the season considered. Each graph is separated in two, for each place (represented on the top part of the graph).

2.4.4 Complete data set

Having analysed all possible combinations from this data set, the complete trajectories of the biological variables can now be addressed. In Figure 2.11, the plots for *total abundance*, *taxonomic richness* and *biotic coefficient* can be found. In these plots, the trends identified in the previous sections become more obvious, particularly the decrease in organisms and taxa presented in 2007-2008. The effect of seasonality in these data is quite subtle but, nevertheless, important. This effect is observed in each year and across all years, where a fluctuation of values between winter, spring, summer and autumn is quite abrupt sometimes. There is another consideration that must be made about this data. For some sampling stations, particularly in *Portinho da Costa*, some extreme values can be observed in 2008 and 2010.

Besides this, a few simple GLMs were created in order to understand the relationship between each one of the biological and substrate variables. With these results, it was possible to conclude that there is no evidence of such direct relation for the study here presented. This conclusion is possible due to the fact that, with the available data, no significance was found for the parameters associated with any of the explanatory variables.

Considering the findings above mentioned, the main analysis of this data will be conducted with some caution.

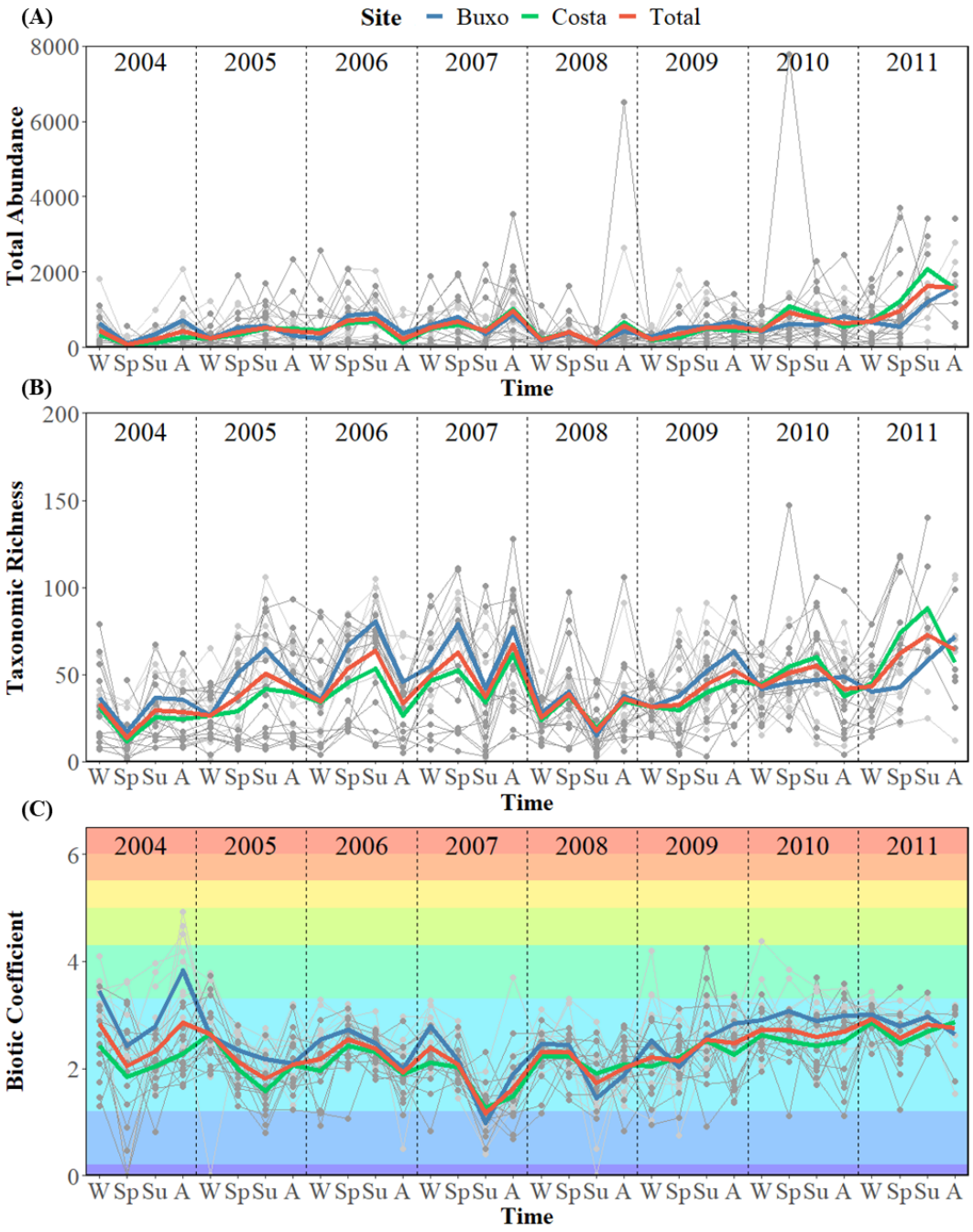

Figure 2.11: Plot of the observed values from the three biological variables in all the 24 sampled stations (light grey for the samples collected in *Porto do Buxo* and a darker grey for the samples collected in *Portinho da Costa*) with the mean trajectory for each one of the two sites (blue for *Porto do Buxo* and green for *Portinho da Costa*) and the mean trajectory considering all data (red). Separation by dashed lines represents the year division of the samples with the respective year identified on top. (A) Total Abundance; (B) - Taxonomic Richness; (C) - Biotic Coefficient

3 | Bayesian Methodology and Latent Growth Curve Models

In this chapter, the basic ideas of Bayesian statistics will be presented, as well as the main differences between Bayesian and classical inference. Description and methodologies to analyse longitudinal data will also be referred here. This chapter intends to provide the reader with some important tools to fully interpret the next chapters.

Before knowing how Bayesian statistical inference works, a simple knowledge about the differences between Bayesian and classical/frequentist statistical inference is needed [45, 46]. The two statistical approaches mentioned, differ in four main aspects:

- **Main goal**. While classical statistics estimates the probability of data conditional on a hypothesis ($p(D \mid H)$), Bayesian statistics focuses on quantifying the probability of some hypotheses being true in the light of our data ($p(H \mid D)$);
- **Definition of probability**. In classical statistics, probability is defined in terms of the limit of the relative frequency of an event, while in Bayesian statistics, probability is defined as the "degree of belief" in the likelihood of an event and can be different depending on the user's knowledge of the subject;
- **Previous information**. In Bayesian statistics, parameters' prior information is used to complement the observed data, whereas in classical statistics only data are used and previous information about the parameters is ignored;
- **Model parameter status**. Classical statistics treats model parameters as fixed, "true" quantities whereas in Bayesian statistics they are considered as random variables.

Bayesian statistics has several advantages over classical, being the most important the ability to modify beliefs we have about the model parameters based on the observed data and not on all possible data sets that might have occurred but did not. Considering the parameters as random variables enables the usage of probabilistic statements about them. The Bayesian paradigm focuses on using probabilities to express acquired knowledge and combining those with the results of new experiments to update this knowledge. This is done by asking what is the probability of a certain hypothesis, H_i, from a set of p hypothesis $(H_1, H_2, \ldots, H_p)$, being true given that a data set, D, was observed [47, 48]. The updated information regarding H_i after data D is observed, is given by Bayes' theorem [46].

$$p(H_i \mid D) = \frac{p(H_i \cap D)}{p(D)} = \frac{p(D \mid H_i) \times \pi(H_i)}{p(D)}, p(D) > 0 \tag{3.1}$$

where $p(D \mid H_i)$ is known as the likelihood function conditional on the hypotheses H_i (can also be represented as $L(H_i \mid D)$), $p(H_i)$ is known as the prior distribution (also represented as $\pi(H_i)$), and reflects the information previous to the conduction of the experiment. $p(H_i \mid D)$ is known as the posterior distribution and reflects the updated results. $p(D)$ is the normalizing constant that represents the marginal density across all possible hypothesis and does not depend on i, *i.e.*, it is not associated with the hypothesis that is being considered at the moment. Therefore, the posterior probability of a hypothesis H_i given D is proportional to the product of their prior probabilities and likelihoods, as:

$$p(H_i \mid D) \propto L(H_i \mid D) \times \pi(H_i) \tag{3.2}$$

3.1 Bayes' Theorem

Suppose that $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$ is a set of mutually exclusive parameters ($p(\theta_i \cap \theta_j) = 0, i \neq j$) and at least in some point in time, exhaustive. The latter assumption is very important because it is impossible to formulate all parameters at any given time, but they become a possibility when considering a single point in time, *i.e.*, the parameters can only be addressed based on the knowledge one has at the moment [47]. From this set of parameters, the "true" one cannot be observed, however, one may accept the most likely to occur. This means that with m parameters, one must have a higher likelihood of being true when compared to other alternatives.

Let $p(\theta_i)$ be the assigned probability to the i-th parameter. By the probabilities' properties we have:

$$0 \leq p(\theta_i) \leq 1, i = 1, 2, \ldots, m$$

$$\sum_{i=1}^{m} p(\theta_i) = 1$$

Thus, $p(\theta_i)$ is the prior probability assigned to θ_i when the set of m hypothesis is competing. Having defined the parameters' prior distribution, let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ be a set of n data points obtained from an experiment. Given that θ_i is considered to be true, these results should occur with conditional probabilities:

$$0 \leq p(x_j \mid \theta_i) \leq 1, i = 1, 2, \ldots, m, j = 1, 2, \ldots, n$$

$$\sum_{j=1}^{n} p(x_j \mid \theta_i) = 1$$

This gives the conditional probability of x_j being observed under θ_i. Then let X be a random variable that can take one of the n possible values x_j ($j = 1, 2, \ldots, n$), such that $\boldsymbol{\theta}$ and X have a joint distribution given by:

$$p(\boldsymbol{\theta} = \theta_i, X = x_j) = p(x_j \mid \theta_i) \times p(\theta_i)$$

As stated before, in Bayesian statistics the main objective is not to evaluate how the data could have been observed, if a certain hypothesis is true, but to determine what is the probability of some hypothesis, θ_i, being true given a datum, x_j ($p(\theta_i \mid x_j)$).

$$p(\theta_i \mid x_j) = \frac{p(\boldsymbol{\theta} = \theta_i, X = x_j)}{p(x_j)} = \frac{p(x_j \mid \theta_i) \times p(\theta_i)}{p(x_j)} \tag{3.3}$$

where $p(x_j)$ is the probability of observing x_j throughout all possible hypothesis and can be determined using the Law of Total Probability:

$$p(x_j) = p(x_j \mid \theta_1) \times p(\theta_1) + p(x_j \mid \theta_2) \times p(\theta_2) + \cdots + p(x_j \mid \theta_m) \times p(\theta_m) = \sum_{i=1}^{m} p(x_j \mid \theta_i) \times p(\theta_i)$$

Thus, replacing this term in Equation (3.3):

$$p(\theta_i \mid x_j) = \frac{p(\mathbf{x}_j \mid \theta_i) \times p(\theta_i)}{\sum_{i=1}^{n} p(x_j \mid \theta_i) \times p(\theta_i)} \tag{3.4}$$

This is known as the application of Bayes' Theorem to a cause-effect problem, *i.e.*, identifying the most plausible cause from the effects observed. Equation (3.4) illustrates the concept of Bayesian learning, the process by which a prior opinion is changed by evidence and becomes the posterior one. Supposing that, after experiment 1, that produced the data set $x^1 = (x_1^1, x_2^1, \ldots, x_{n_1}^1)$, a new experiment is conducted, producing a new data set, $x^2 = (x_1^2, x_2^2, \ldots, x_{n_2}^2)$. By Bayes' theorem, the new posterior is given by:

$$\begin{aligned} p(\theta_i \mid x_{j^2}^2, x_{j^1}^1) &= \frac{p(x_{j^2}^2, x_{j^1}^1 \mid \theta_i) \times p(\theta_i)}{\sum_{j^2=1}^{n_2} p(x_{j^2}^2, x_{j^1}^1 \mid \theta_i) \times p(\theta_i)} \\ &\propto p(x_{j^2}^2, x_{j^1}^1 \mid \theta_i) \times p(\theta_i) \\ &\propto p(x_{j^2}^2 \mid x_{j^1}^1, \theta_i) \times p(x_{j^1}^1 \mid \theta_i) \times p(\theta_i) \\ &\propto p(x_{j^2}^2 \mid x_{j^1}^1, \theta_i) \times p(\theta_i \mid x_{j^1}^1) \end{aligned} \tag{3.5}$$

The preceding equation shows that the posterior distribution for θ_i given the two data sets will be proportional to the likelihood associated to x^2 and the "new" prior information of θ_i. In this case, before experiment 2, there is information available from the first experiment conducted, so the prior distribution will be equal to the posterior distribution of θ_i after performing experiment 1. It is worth noting that, given θ_i, $x_{j^1}^1$ and $x_{j^2}^2$ ($j^1 = 1, 2, \ldots, n_1, j^2 = 1, 2, \ldots, n_2$) are conditionally independent, which implies that,

$$p(x_{j^2}^2, x_{j^1}^1 \mid \theta_i) = p(x_{j^2}^2 \mid x_{j^1}^1, \theta_i) \times p(x_{j^1}^1 \mid \theta_i) = p(x_{j^2}^2 \mid \theta_i) \times p(x_{j^1}^1 \mid \theta_i)$$

Continuous case

Many models currently used are continuous and so is the parameter space. For this reason, a simple revision must be done for this type of models. In cases where the parameters and/or data are continuously distributed, Bayesian process is essentially identical to the discrete case, having only slight differences. Consider that both $\mathbf{x}$ and $\boldsymbol{\theta}$ are assumed to be able to take any continuous value contained in their spaces defined as sampling space ($\mathfrak{R}_{\mathbf{x}}$) and parameter space ($\boldsymbol{\Theta} = \{\theta : p(\theta) > 0\}$), respectively [48].

Consider $\pi(\boldsymbol{\theta})$ as the prior distribution attributed to the parameter vector and $f(\mathbf{x} \mid \boldsymbol{\theta})$ as the probability density function of $\mathbf{x}$ given $\boldsymbol{\theta}$. Given that $(x_1, x_2, \ldots, x_n)$ is a realization of a random sample, then:

$$f(\mathbf{x} \mid \boldsymbol{\theta}) = f(x_1 \mid \boldsymbol{\theta}) \times f(x_2 \mid \boldsymbol{\theta}) \times \cdots \times f(x_n \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i \mid \boldsymbol{\theta}) \tag{3.6}$$

Thus, using Equation (3.1), the posterior distribution of $\boldsymbol{\theta}$ given the observations $\mathbf{x}$ is defined as:

$$p(\boldsymbol{\theta} \mid \mathbf{x}) = \frac{\pi(\boldsymbol{\theta}) \times f(\mathbf{x} \mid \boldsymbol{\theta})}{\int_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \pi(\boldsymbol{\theta}) \times f(\mathbf{x} \mid \boldsymbol{\theta}) d\boldsymbol{\theta}} \tag{3.7}$$

The posterior distribution is frequently impossible to derive analytically or the calculations are to complicated and time consuming. To overcome this problem, computational methods were developed, being the most used, the Markov Chain via Monte Carlo (MCMC) simulation. This method will be summarised in section 3.4.

3.2 Prior distributions

The selection of a prior distribution is an important step in Bayesian inference, as it represents the knowledge about unknown parameters before seeing the data. Therefore, it is useful to understand the types of existing prior distributions. They can be classified as informative or non-informative and selection between these two types will depend on the knowledge one has about the problem under study, before the experiment is carried out.

Non-informative priors can be used when there is no information about the parameters (or the available information is too weak), or if an inference based only on the data is desired, *i.e.*, the objective is to "let the data speak". Some common choices of non-informative prior distributions are:

- **Uniform prior** can be used both for discrete and continuous cases. It assumes that all values from the parameter space are equally likely. For instance, in the discrete case, for the parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$:

  $$p(\theta_i) = \frac{1}{m}, i = 1, 2, \ldots, m$$

  For the continuous case, the continuous Uniform prior distribution is used.

- **High variance prior** are the ones that may use a known distribution to describe the parameters but considers a large variation, resulting in a vague distribution. For instance, consider a $Gamma(\alpha, \beta)$ $\alpha, \beta > 0$, where α and β are very small (for example 0.001). The result is a distribution with a large variance that is commonly used to express weak knowledge, e.g., in the case of a scale parameter;

- **Jeffreys' prior** uses the Fisher Information, $I(\theta)$, of a model to determine a non-informative prior distribution for the parameters [49]. Consider $f(x \mid \theta)$ to be the probability density function of X, where $\theta \in \Theta$ is a scalar. Assuming that $log(f(x \mid \theta))$ is twice differentiable in θ, Fisher Information is obtained as follows:

  $$I(\theta) = -E_{X|\theta}\left[\frac{\partial^2}{\partial \theta^2} log(f(x \mid \theta))\right]$$

  Jeffreys' prior is given by:

  $$\pi_J(\theta) \propto [I(\theta)]^{\frac{1}{2}}$$

Informative priors are associated with Bayesian inference's subjectivity, because different priors can lead to different results. For this reason, the choice of *prior* to use in the analysis is a very important issue [50].

3.3 Bayesian inference

3.3.1 Inference on posterior distributions

Once the *posterior* is determined it can be used to make inferences about the model parameter(s). If possible, it is most convenient to do so by using a graphical representation. Once the *posterior* is available, point estimations can be obtained such as the posterior mean and variance, as well as credible regions.

Given the posterior distribution, the mean can be calculated as:

$$E[\theta \mid \mathbf{x}] = \int_{\Theta} \theta p(\theta \mid \mathbf{x}) d\theta$$

while, the variance is given by:

$$Var[\theta \mid \mathbf{x}] = \int_{\Theta} (\theta - E[\theta \mid \mathbf{x}])^2 p(\theta \mid \mathbf{x}) d\theta$$

In classical statistics, confidence intervals (CI) are used to indicate a range of values which, if the experiment that generated the current data is repeated many times, $100 \times (1-\alpha)\%$ of the intervals obtained will contain the true parameter value, whereas the remaining $100 \times \alpha\%$ will not. In Bayesian statistics there is an analogous concept known as credible region of θ. These regions can be determined using the posterior distribution obtained from the observed data (not from data that may never be observed). The region with $100 \times (1-\alpha)\%$ probability of containing the true value of the parameter *a posteriori* is defined as follows:

$$1 - \alpha \leq \int_{L}^{U} p(\theta \mid \mathbf{x}) d\theta$$

where L and U represent the lower and upper bounds of the $100(1-\alpha)\%$ credible region for θ.

Given the posterior distribution of θ and a significance value (α), many credible regions can be obtained for a region of $100 \times (1-\alpha)\%$ of credibility. One region that is commonly determined is the equal tail probability interval, for which the significance value is divided in half and the quantiles $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ of the posterior distribution are determined. This may be an alternative when the posterior distribution is unimodal and symmetrical. In the cases where it is not so, another interval can be obtained, the Highest Posterior Density (HPD) interval. HPD credible region of θ (I_α) is given by:

$$I_\alpha = \{\theta : p(\theta \mid \mathbf{x}) \geq \gamma\}$$

where γ is a value such that:

$$P(\theta \in I_\alpha \mid \mathbf{x}) = 1 - \alpha$$

3.3.2 Predictive distribution

In data modelling, the main objective is, in general, to predict future data values. In Bayesian inference, this process is done by means of a distribution, which is based on the distribution of a future observation, $\mathbf{y}$, given the parameter, θ and the data, $\mathbf{x}$, $f(\mathbf{y} \mid \theta, \mathbf{x})$, and the updated information about θ existing at the moment, *i.e.* the posterior distribution, $p(\theta \mid \mathbf{x})$. This new distribution is called predictive distribution, $m(y \mid \mathbf{x})$, of a new observation and can be determined using Equation (3.8).

$$m(y \mid \mathbf{x}) = \int_{\Theta} f(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{x}) p(\boldsymbol{\theta} \mid \mathbf{x}) d\boldsymbol{\theta} \tag{3.8}$$

This equation can be simplified when the new observation, y, and the data, $\mathbf{x}$, are conditionally independent given $\boldsymbol{\theta}$.

$$m(y \mid \mathbf{x}) = \int_{\Theta} f(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{x}) d\boldsymbol{\theta} \tag{3.9}$$

3.4 Markov Chains and the MCMC process

Without loss of generality, let the sequence of random variables $X_n (n = 0, 1, 2, \dots)$ be a finite state, discrete Markov Chain where X_n can take values in a set $S = \{0, 1, \dots, n-1\}$, called the states of the Markov Chain. Subscripts n are called stages or time periods, so if $X_n = i$, the process is in state i at time n. This process must satisfy the Markov property stated below that establishes the relationship between Markov Chain states using the transition probability:

$$p(X_n = j \mid X_{n-1} = i, \dots, X_0 = m) = p(X_n = j \mid X_{n-1} = i) = p(j \mid i), i, j, \dots, m \in S$$

which means that, in a Markov Chain, the conditional distribution at time n, given the past states $X_{n-1} = i, \dots, X_0 = m$, only depends on the immediately preceding one observed, $X_{n-1} = i$. The transition probability, $p(j \mid i)$, describes the evolution of the Markov Chain and represents the probability that, in time n, the chain is in state j given that in the preceding time, $n-1$, it was in state i, being a conditional probability where j is stochastic and i is fixed.

Considering that the n states can be enumerated, the transition probabilities can be arranged in a $n \times n$ matrix $\mathbf{P} = \{p(j \mid i)\}$, where rows are represented by i and columns by j. Consider that $\mathbf{P}$ is independent from time or is associated to a time homogeneous chain, *i.e.*, the probability of transition between any two states depends only on the states and not on the time period that the chain is on. The $(i+1)$-th row of $\mathbf{P}$ represents the probability distribution of X_n given that $X_{n-1} = i$. As the entries of this matrix are probabilities, they follow the usual probability properties: $0 < p(j \mid i) < 1$ and in each row $\sum_j p(j \mid i) = 1$.

MCMC has become a very important method to draw inferences from complex posterior distributions, using computational methods when analytical expressions are difficult to handle. This process is based on the idea of generating Markov Chains via Monte Carlo simulation that has asymptotically, the desired posterior distribution as its equilibrium or stationary distribution.

There is a main algorithm developed for the implementation of MCMC methods, the Metropolis-Hastings [51, 52], of which the Gibbs sampler [53] is a particular case. Taking into account that many real problems involve continuous variables, it is worth noting that, in this particular case, the usage of a transition matrix is not applicable, so transition kernels are used instead (see [54] for more details).

3.5 Longitudinal data and the Latent Growth Curve Models

In many scientific areas, from clinical trials to environmental studies, studying time-related variables is a major aspect, leading to repeated measurements on the same individuals throughout time, generating what is usually called longitudinal data [55, 56]. Collecting this sort of information can have many purposes, such as inference of trait trajectories within individuals and populations or even assess variation sources between and within groups. Although this type of data can be difficult to analyse, many studies

concluded that this is a necessary step to understand developmental processes [57]. There are several different methods to analyse this type of data, particularly by using latent growth curve models, which have been identified as an important research technique [58]. This method is mainly used to describe, test, evaluate and make inferences of an organism's trajectory throughout time, focusing not only on the growth part of the process, but also on the decline, oscillation and stabilization of the systems.

The application of these models can be a complex process. Therefore McArdle & Grimm (2010) [57, 59] proposed a five-step method to ease this task. From five steps proposed, the first three are the most important to understand and describe the development of single variables without the influence of others in order to compare sites in terms of trends. These steps are as follows:

1. **Data description**: using trajectory plots to create an overview of the data set, emphasizing the most important aspects detected;

2. **Characterization of subjects and groups' trajectories**: describing experimental units and groups' characteristics involved in the study, representing them by models;

3. **Examine inter-individual differences in developmental shapes**: with trajectories description and modelling, variability source existing on an individual level has to be assessed.

3.5.1 Latent Growth Curve Model

For each individual i ($i = 1,2,\ldots,N$), trajectory characterization is done by considering one of the variables of interest. First step is to write the trajectory equation in a mixed-model form (see Equation (3.10)). This equation considers observed values (y_{it}) for the $i-th$ individual at time t ($t = 1,2,\ldots,T$, where T is the number of times each individual is observed) and three latent or unobserved variables, the random-effects parameters, initial level (L_i), initial slope (S_i) and a random error (e_{it}). To define the trajectory's shape over time, another parameter is added to the model specification (α_t), known as the shape parameter. Since initial level and slope of an individual are traditionally allowed to co-vary, the formulation of these two parameters can be done bivariately. To do so, a new variable is presented as a person-specific parameter (LS).

$$\begin{aligned} y_{it} &= L_i + \alpha_t S_i + e_{it} \\ (LS)_i &\sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ e_{it} &\sim N(0, \sigma_e^2) \end{aligned} \tag{3.10}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are defined as follows:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_L \\ \mu_S \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_L^2 & \sigma_{LS} \\ \sigma_{LS} & \sigma_S^2 \end{bmatrix}$$

The latent initial level of an individual i, L_i, is defined by a mean latent level (μ_L), which is equal to all of them and the initial variance around this value (σ_L^2) representing the differences between individuals, that is:

$$L_i = \mu_L + v_{L_i}$$

The same kind of model is applied to the initial slope of the $i-th$ individual, which is defined by a

common mean value (μ_S), again having the same value for all individuals and a variance term representing the variability between individuals around the initial slope S_i (σ_S^2), that is:

$$S_i = \mu_S + v_{S_i}$$

To control how these two parameters interact, the covariance (σ_{LS}) between them is also taken into account.

Considering that the model is implemented from a Bayesian point of view, parameters μ_L, μ_S, σ_L^2, σ_S^2, σ_{LS} (through $\mathbf{T}$, precision matrix), σ_e^2 (through τ_e, the precision) and α_t are random quantities. Their uncertainty has to be introduced in the model using prior distributions. Following Oravecz *et al.* (2018) [60] the prior distributions considered were:

$$\begin{aligned}
\mu_L &\sim N(a, b^2) \qquad \mu_S \sim N(c, d^2) \\
\mathbf{T} &\sim Wishart(\mathbf{E}_{2x2}, n), n > 1 \\
\tau_e &\sim Gamma(f, g), f > 0, g > 0 \\
\alpha_t &\sim N(h_t, i_t^2), t = 1, 2, \ldots, T
\end{aligned}$$

where a, b, c, d, $\mathbf{E}$, f, g, h_t and i_t are the hyperparameters.

Figure 3.1 represents the path diagram for a growth curve model and creates a simpler way to visualize the process. Three different types of variables present in this process are represented with different shapes, being the triangle a constant, the circles latent or unobserved variables and the squares the observed data points. The diagram represented below indicates that the observations are directly dependent on individual initial level (L), initial slope (S) and random errors (e), having L and e a direct relation (mediated by a factor of 1) and S mediated by the shape parameter α_t. This mediation by α_t is determinant to trajectory analysis. In order to decrease the high autocorrelation present between these parameters, at least two of the t possible values of $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_t)$ have to by priorly fixed.

3.5.2 Inter-group variability analysis

With latent growth curve models implemented, results can be used to plot the overall trajectory over time. Also, estimated values of α_t allow to make inferences about differences at each point of time. Although, through the analysis of the results obtained by the latent growth curve model, a lot of information can be obtained about the process variability, there is no way to directly analyse the source of it; only its presence or absence can be identified. One way to proceed after this first approach is to apply new techniques, that allow the evaluation of differences' source. Several methods exist to do so and, in this work, multiple group latent growth curve model will be used.

Let X be a dichotomous random variable that indicates a difference between individuals (*e.g.* gender, sampling centre, age class). Without loss of generality, assume that there are two different groups in the data. Considering one of them as reference, the variable X will take the value 0 if the $i-th$ individual belongs to this reference group and it will take the value 1 otherwise.

Considering that a grouping variable like X exists, a LGCM with grouping factors can be implemented. In general, the two models are similar, having a slight difference when it comes to the individuals' initial level and slope. Again, considering that there are two groups (g_1 and g_2) in the study and that group g_1 is the reference, the relations

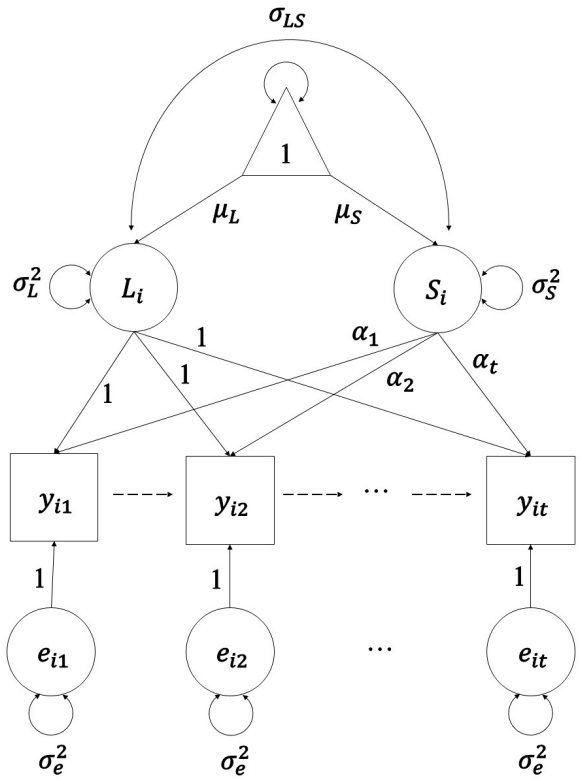

Figure 3.1: Adapted from Zhang *et. al* (2007) [55]. Growth curve model path diagram where the triangle represents a constant, the circles represent latent unobserved variables (individual level, slope and random error) and the squares the observed values. Dashed arrows represent the time relation between observations from the same experimental unit/individual.

$$\mu_{L_i} = \mu_{L_{g_1}} + \mu_{L_{g_2}} X_i \tag{3.11}$$

$$\mu_{S_i} = \mu_{S_{g_1}} + \mu_{S_{g_2}} X_i \tag{3.12}$$

$$\alpha_t = \alpha_{t_{g_1}} + \alpha_{t_{g_2}} X_i \tag{3.13}$$

are obtained.

With this new formulation of the LGCM, it is possible to assess the assumption of invariance made in the first model, *i.e*, in the first model, all groups in the study are considered to have the same variability/trajectory over time and in this "new" model, each group is assigned with its' own initial level, slope and shape. L_{g_1}, L_{g_2}, S_{g_1}, S_{g_2}, $\alpha_{t_{g_1}}$ and $\alpha_{t_{g_2}}$ are then the parameters for each one of the two groups (g_1 and g_2). In what concerns the prior distributions for these new parameters, the same ones chosen before for μ_L, μ_S and α could be used.

4 Application of the LGCMs to Ecological Monitoring data

In this chapter, the results obtained from the application of the methods described previously are presented. According to the work of McArdle & Grimm (2010) [57], Zhang *et. al* (2007) [55] and Oravecz *et. al* (2018) [60], this chapter is divided into three parts. First, a brief description of the models used to obtain the results is presented. Secondly, a simple description of trajectory plots and statistics in order to identify trends or major events follows. Finally, the results obtained from LGCM's application to each variable are presented and briefly analysed.

Considering the findings presented in Chapter 2 about the data and the relations found, the analysis through the LGCMs will be done using a subset of the initial data set. In this analysis, only the last four years (2008 to 2011) of sampling will be used to avoid the interference of the inexplicable event that occurred in 2007-2008. From this subset of the original data, seasons will be separated to eliminate the seasonal effect on the biological variables. Given this new cut and re-arrangement of the data set, there are 24 sampling stations repeatedly measured 16 times. Using the season as a separation variable, the final data set is comprised of 4 observations for each sampling station (one for each year) and a total of 4 models will be created for each biological variable (one for each season).

4.1 The models

As formerly described in Chapter 2, the distributions used to model the three variables of interest were the Negative Binomial for *Total Abundance* and *Taxonomic Richness* and the Normal distribution for the *Biotic Coefficient*. For more details on models' formulation, adapted from Zhang *et. al* (2007) [55], see Appendix C. Given these choices of distributions for the data, some explanations are needed about the meaning of the parameters. To model count data, the Poisson distribution is commonly referred as the most indicated. The main issue with this distribution is that the expected value and variance are equal. This assumption makes it difficult to deal with overdispersed data. To accommodate this overdispersion, the Negative Binomial distribution can be used. Considering the later, there are two formulations used to define a random variable Y which has a Negative Binomial distribution, one using the probability of success, p, and r; and another that uses the mean value, μ, and r. The random variable Y counts the number of failures in a sequence of independent Bernoulli trials before r successes (with probability p) are obtained. This distribution has mean value of μ and variance of $\mu + \mu^2/r$. Consequently, the variance of this distribution is increased, relatively to the variance of a Poisson distribution, by the amount μ^2/r. As such, this distribution can model data which exhibits an overdispersion feature. In Equation (4.1), the probability mass function of a Negative Binomial (r, μ) is presented:

$$f(y \mid \mu, r) = \frac{\Gamma(y+r)}{y!\Gamma(r)} \times \frac{r^r \mu^y}{(r+\mu)^{r+y}}, y = 0, 1, 2, \ldots \tag{4.1}$$

Taking into account all that was described in the previous chapter on LGCMs, a short explanation of the prior distributions chosen, as well as the likelihood functions used for the data modelling, is necessary to proceed successfully with the analysis of the results. In Table 4.1 it is possible to find all the prior distributions used in this work. They are presented in a similar format as in JAGS, meaning that, in the Normal distribution, the second parameter considered is the precision, τ ($\tau = 1/\sigma^2$), and not the variance, σ^2. Considering that there was no previous information about the parameters, the prior distributions chosen were all non-informative.

Table 4.1: Prior distributions used in the models for latent parameters estimation.

	Non-informative Prior
α_2, α_3	$dnorm(0, 0.001)$
μ_L	$dnorm(0, 0.001)$
μ_S	$dnorm(0, 0.001)$
r	$dunif(0, 50)$
$1/\sigma_e^2$	$dgamma(0.01, 0.01)$
$\begin{bmatrix} \sigma_L^2 & \sigma_{LS} \\ \sigma_{LS} & \sigma_S^2 \end{bmatrix}$	$dwish\left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, 2\right)$

In principle, there is no restriction on the so-called shape parameters, α_2 and α_3. Based on the data from 2004 up to 2007, very large (small) changes in the slope, in the overall, were not expected. Therefore, the traditional vague normal distribution was used based on some indications of several other authors [50, 55].

4.2 Data description

In this work, the experimental units are the sampling stations, meaning that the data set contains 24 experimental units, measured repeatedly 16 times over 4 years. Having in mind Figure 4.1, regarding the Winter, in all three variables, a linear trend can be observed with values increasing from 2008 to 2011. In Spring, there is also a growing trend, although, from 2008 to 2009, there are no strong visual differences with regard to the organisms' abundance and diversity. For the communities' health there is a slight improvement in the same time frame. For Summer, the same growing linear trend observed for Winter is present but with a higher slope, indicating that in this season there is a tendency to have communities with more organisms and taxa. Finally, in Autumn, although there is an overall growing trend between 2008 and 2010, there is a stationary pattern in terms of organisms' abundance and taxonomic richness.

Observing all data points of the three biological variables under study (grey bullet points in Figure 4.1), it is worth noting that there is a high variability in observations from certain seasons like Spring and Autumn. This is a natural variability given that biological systems are not stable and different species tend to respond differently to environmental variations.

Being aware that, in previous analysis, there were some differences in the two sampled sites, the modelling process was conducted considering all station without the site division. This decision was made based on the small number of experimental units available in each site and on the constraints of the methods. In the data set, there are 9 experimental units in *Porto do Buxo* and 15 in *Portinho da Costa*. Although this difference might not be significant in ecological terms, to apply LGMs for this reduced

amount of units might lead to inaccurate results.

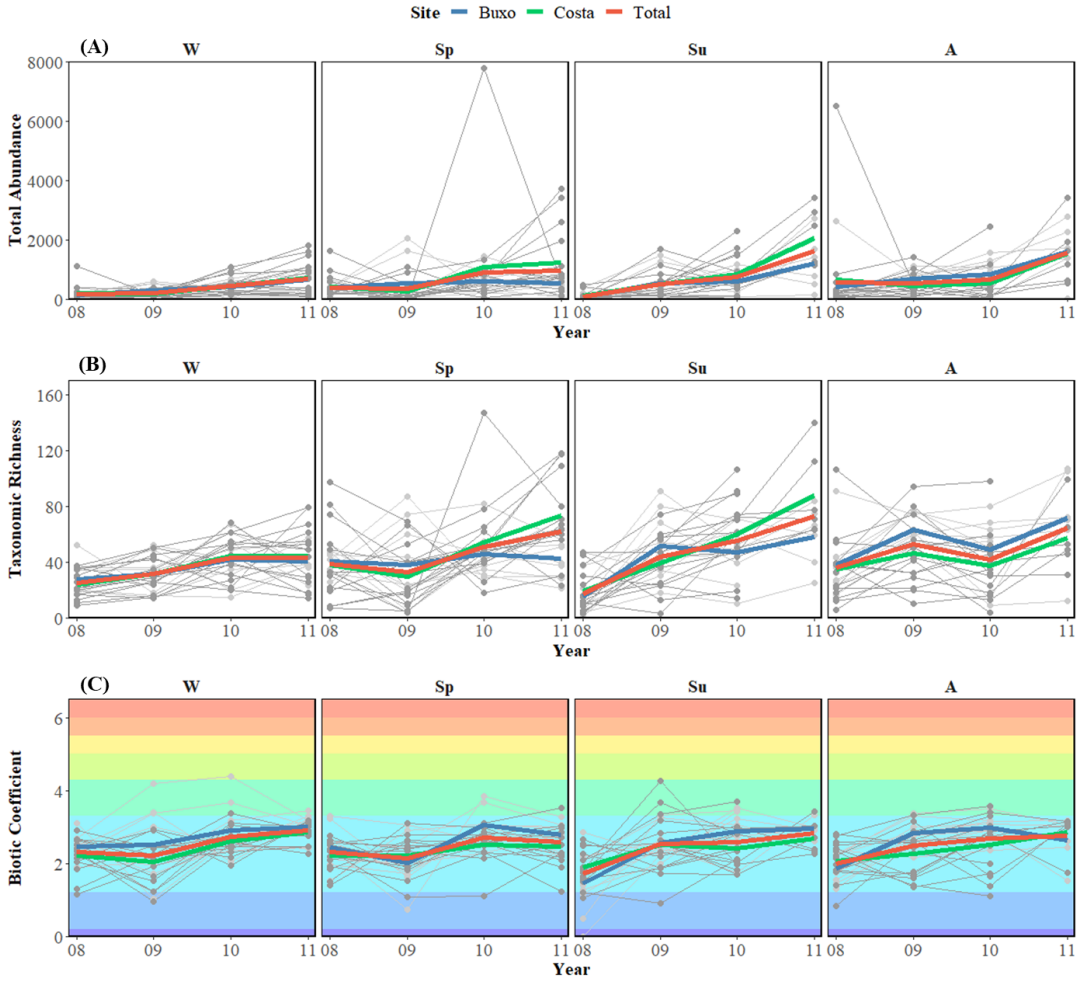

Figure 4.1: Trajectory plot of the observed values from the three variables of interest in all the 24 sampled stations (light grey for the samples collected in *Porto do Buxo* and a darker grey for the samples collected in *Portinho da Costa*) with the mean trajectory for each one of the two sites (blue for *Porto do Buxo* and green for *Portinho da Costa*) and the mean trajectory considering all data (red). Each plot is divided in four, one for each season, identified in the top part of the plots (W - Winter, Sp - Spring, Su - Summer, A - Autumn). (A) Total Abundance; (B) - Taxonomic Richness; (C) - Biotic Coefficient

4.3 Trajectory characterization

For trajectories characterization of the biological variables previously considered, four Bayesian models were created for each variable, one for each season. These models seek to describe the general trajectory of each variable over time. For each model, five Markov chains with a total of 10 thousand iterations each were initiated at random points. From all these chains, the first 1000 iterations (value obtained by trial and error), corresponding to the burn-in period, were removed. In Markov Chains, sampled values are correlated. To deal with this issue, there is a method commonly used called thinning. This procedure consists on the retention of equally spaced sampled values, which reduces not only the total sample size but mainly the association between consecutive values. In this work, a thinning of 50 (value obtained by trial and error) was used, *i.e.*, after removing the burn-in iterations, the remaining values were saved every 50 iterations. The results are composed by five chains with 2 thousand iterations

each.

4.3.1 Convergence Assessment

In order to fully characterize and correctly understand the results of the data modelling, it is essential to perform a brief analysis of the chains obtained by MCMC. This analysis includes a brief evaluation of chains' convergence, using graphical methods and statistical tests and also a model check with the main aim of understanding the models' capability to predict the variables of interest.

Markov Chains' convergence can be assessed using trace plots (Figure 4.2 A and B), since they are able to show the information of thousands of iterations, obtained for each chain, in a simple way. Density plots were also used; an example for μ_L and μ_S can be found in Figure 4.2 C and D. To assess the adequacy of the sample size that was considered, the effective sample size (ESS; number of independent values obtained) was calculated. It is pertinent to note that the larger this value, the smaller the autocorrelation between iterations is. One way to evaluate this autocorrelation, is through the Autocorrelation Function (ACF), which determines the appropriate lag between observations, that is, which interval is necessary for two observations to be approximately independent. To easily perceive the results accomplished by this analysis, the values obtained for the sample autocorrelations are represented graphically (Figure 4.2 E and F). By analysing the six plots, it can be concluded that there is no evidence whatsoever that the chains have not converged.

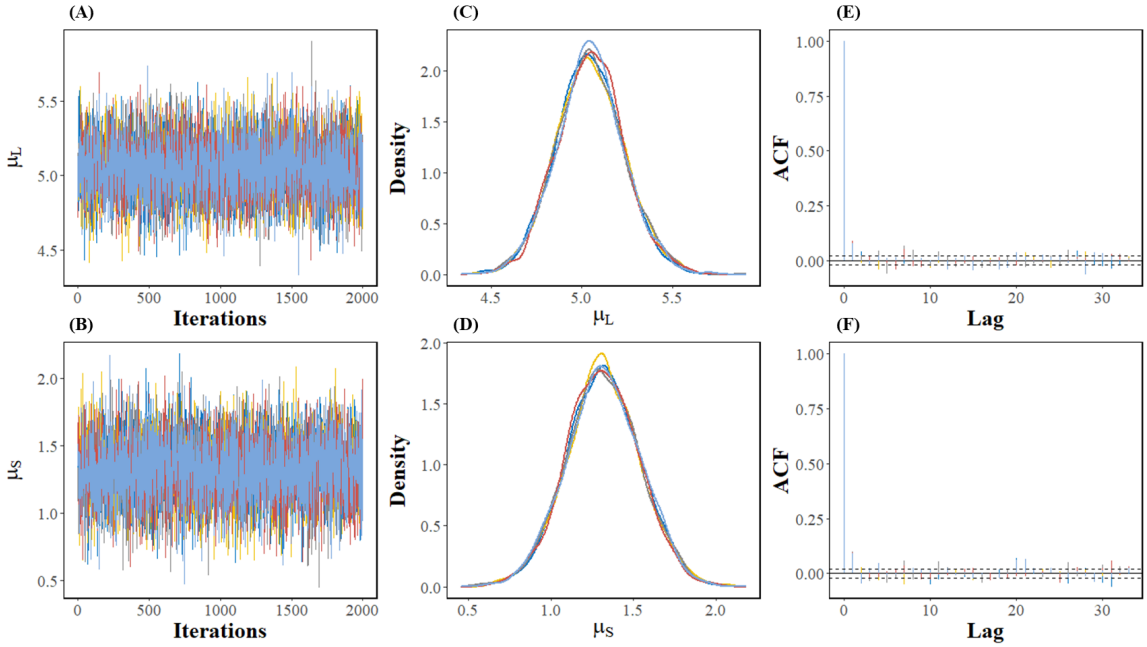

Figure 4.2: Graphical representations used for convergence analysis from two of the parameters modelled (μ_L and μ_S) for *Total Abundance* using data from Winter. (A) and (B) trace plot of the three chains obtained for each parameter; (C) and (D) density plot from the same chains; (E) and (F) autocorrelation function values.

In order to assess the chains' convergence, the $\hat{R}$-statistic is used. $\hat{R}$-statistic corresponds to the ratio of within-and between chain variances. For a certain parameter it is possible to conclude that the chain has converged when the above-mentioned ratio is less than 1.1 [61].

In the previous chapter it was mentioned that fixing two α_t values would decrease the autocorrelation of all the others. Considering the existence of the Latent Basis Growth Model (a type of LGCM), the first and last alphas were fixed while the remaining two were allowed to be freely estimated. Particularly, α_1

was set to 0 and α_4 to 1 for all the models. These choice were based on the information gathered from [50, 58].

Plots like those presented in the Figure 4.2, were obtained for all parameters under study, as well as the descriptive statistics. After a detailed assessment of the results produced it could be concluded that the chains seem to have converged and the results can then be analysed. The remaining plots and full tables can be consulted in Appendix D.

4.3.2 Variable mean description

Model checking is carried out using the observed values of the variables and the estimated values obtained by the models. Considering the problem under study, this assessment consists in comparing observed means of *Total Abundance*, *Taxonomic Richness* and *Biotic Coefficient* with estimated means for all the individuals (sampling stations).

Graphical representations were made to enable the verification of the models fitted to the *Total Abundance* variable and can be observed in Figure 4.3 (A). As shown below, the natural variability of the data collected was generally captured by the models. The same is observed for the remaining two variables, *Taxonomic Richness* and *Biotic Coefficient* (see Figure 4.3 B and C, respectively). Estimated trajectories presented in the figure below were obtained using Equation (3.10) and the mean of the five Markov chains generated for each parameter (values for all parameters can be found in the tables from Appendix D). The trends identified previously in this chapter are also evident in the estimated trends. All these characteristics indicate that the models fitted are appropriate for a complete description of ecological data.

Numerical Summaries - Winter

Posterior statistics of some parameters related to the modelling of *Total Abundance*, are displayed in Table 4.2, including their posterior means, posterior standard deviations, 95 % Equal Tail Probability (ETP) regions and $\hat{R}$-statistics, calculated based on the conjunction of all the chains. The average initial level of *Total Abundance* is 5.047 ($mean(\mu_L)$), translating on an estimate of approximately 156 organisms; the variance for this initial level is 0.384 ($mean(\sigma_L^2)$). Regarding the initial slope ($mean(\mu_S)$) the value obtained was 1.318 (corresponding to an increase of around 425 organisms from the first to the last year), this value is accompanied with a between subject variability of 0.364 ($mean(\sigma_S^2)$). Comparing the values for the posteriors of σ_L^2 and σ_S^2 it can be seen that the between person variability present at the initial slope is larger than the one present in the initial level. This can mean that, although sampling stations began with different values, in terms of total abundance of organisms, the higher variability of the slopes indicates that not all sampling stations have the same increment over time, organisms tend to appear and disappear at different rates in each station. With regard to the posterior estimate of r, the value obtained was 2.946 ($ETP(95\%) = (2.014, 3.959)$), which indicates the presence of overdispersion in the data for this biological variable, as it was expected. Finally, looking at the results obtained for α_2 ($mean(\alpha_2) = 0.202$) and α_3 ($mean(\alpha_3) = 0.712$) one can say that the total abundance of organisms increases from 2008 to 2011. The proportion at which the estimated value is increased or decreased is given by α_t.

Table 4.2: Posterior statistics of the parameters of the models considered for the *Total Abundance*'s model over all sampling stations considering the Winter (sd - Posterior Standard Deviation, ETP - Equal Tail Probability, ESS - Effective Sample Size).

	mean	sd	ETP (2.5%)	ETP (97.5%)	ESS	$\hat{R}$
σ_L^2	0.384	0.173	0.114	0.731	9704	1.000
σ_{LS}	-0.106	0.151	-0.434	0.147	9388	1.000
σ_S^2	0.364	0.212	0.075	0.783	9377	1.000
μ_L	5.047	0.186	4.692	5.426	8908	0.999
μ_S	1.318	0.224	0.881	1.756	8671	0.999
α_2	0.202	0.132	-0.071	0.449	8759	1.000
α_3	0.712	0.113	0.496	0.940	9935	0.999
r	2.946	0.512	2.014	3.959	10000	1.000

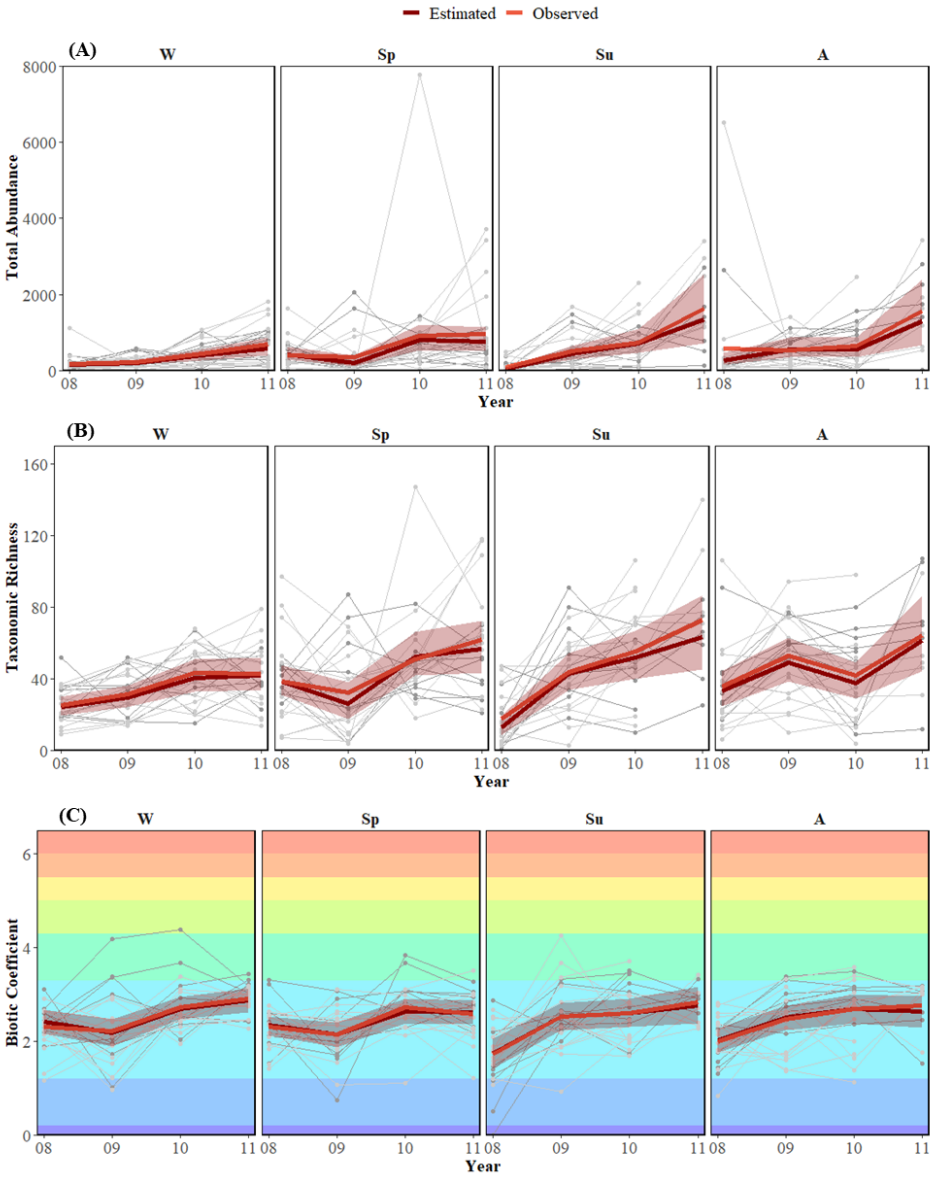

Figure 4.3: Comparison of the estimated means with the observed means for all the sampling stations. Observed values for each sampling station are represented in light grey. Shadowed area represents the 95% Equal Tail Probability region obtained for each model. *Total Abundance* (A); *Taxonomic Richness* (B); and *Biotic Coefficient* (C).

For the initial level of *Taxonomic Richness* (μ_L) a value of 3.184 was obtained, which translates into an average estimate of approximately 24 taxa. As seen for *Total Abundance*, *Taxonomic Richness* also presented a rather low slope ($mean(\mu_S) = 0.547$, corresponding to an increase on 18 taxa from the first to the last year). Comparing the obtained posterior values for the variances of L and S (σ_L^2 and σ_S^2), it is possible to conclude that the individual variability of the initial level and slope are similar (see Table 4.3). For the parameter r, the value 13.384 was obtained ($ETP(95\%) = (7.332, 20.286)$), indicating that, as for *Total Abundance*, overdispersion is present.

Table 4.3: Posterior statistics of the parameters of the models considered for the *Taxonomic Richness*' model over all sampling stations considering the Winter (sd - Posterior Standard Deviation, ETP - Equal Tail Probability, ESS - Effective Sample Size).

	mean	**sd**	**ETP (2.5%)**	**ETP (97.5%)**	**ESS**	$\hat{R}$
σ_L^2	0.165	0.067	0.056	0.295	10000	1.000
σ_{LS}	-0.076	0.061	-0.199	0.025	10000	1.000
σ_S^2	0.187	0.085	0.064	0.353	10000	1.000
μ_L	3.184	0.109	2.974	3.397	10000	0.999
μ_S	0.547	0.128	0.297	0.798	9438	0.999
α_2	0.369	0.151	0.080	0.668	10000	1.000
α_3	0.948	0.153	0.667	1.263	10000	0.999
r	13.384	3.506	7.332	20.286	10000	0.999

Table 4.4 shows that the estimate of the initial level obtained for the *Biotic Coefficient* is 2.421 ($mean(\mu_L)$) points on the previously specified scale (see Table 2.3), with between subject variability, $mean(\sigma_L^2)$, of 0.163. Given the small variation in the observed values, the estimated slope is small ($mean(\mu_S) = 0.451$) as expected. The between subject variability of the slope is higher than the between subject variability of the initial level (σ_S^2 and σ_L^2, respectively) as it was observed for the the *Total Abundance*. It can be said that the biotic coefficient has increased in the last four years of the study, which is related to a decline in the ecological status of the communities. Regarding the variance of the random errors ($mean(\sigma_e^2)$), the value 0.232 was obtained ($ETP(95\%) = (0.152, 0.321)$).

Table 4.4: Posterior statistics of the parameters of the models considered for the *Biotic Coefficient*'s model over all sampling stations considering the Winter (sd - Posterior Standard Deviation, ETP - Equal Tail Probability, ESS - Effective Sample Size).

	mean	**sd**	**ETP (2.5%)**	**ETP (97.5%)**	**ESS**	$\hat{R}$
σ_L^2	0.163	0.066	0.064	0.294	9744	1.000
σ_{LS}	-0.069	0.060	-0.192	0.031	10000	0.999
σ_S^2	0.193	0.089	0.066	0.369	10000	0.999
μ_L	2.021	0.127	1.776	2.267	10000	0.999
μ_S	0.451	0.166	0.141	0.773	10000	0.999
α_2	-0.649	0.465	-1.591	0.163	10000	1.000
α_3	0.615	0.261	0.085	1.117	10000	1.000
σ_e^2	0.232	0.045	0.152	0.321	10000	0.999

In what concerns the covariance between the initial level and slope (σ_{LS}) for *Total Abundance*, the value is higher than the ones obtained for *Taxonomic Richness* and *Biotic Coefficient*, although all estimated values are close to zero. This means that the slope of the trajectories, in all variables, is not related to the initial level. With some knowledge about the relationship of these three variables, it is possible to say that, although the community has been growing in numbers, both in terms of organisms and taxa, the sensitivity of these animals is higher to organic enrichment. This conclusion is based on the

increasing trend observed for the *Biotic Coefficient*.

The tables presented and comments done so far focused only on the models created for the Winter season data. The main differences of the results obtained by the models fitted to the other three season data will be presented next.

Numerical Summaries - Spring

A similar behaviour to the one identified for the Winter is observed for the variable *Total Abundance* during Spring. The main change in the estimated values obtained for α_2 and α_3 since a negative sign appears for α_2 ($mean(\alpha_2) = -1.407$) indicating that, from 2008 to 2009, the total abundance of organisms decreased. The same appears for the *Taxonomic Richness* in the same time period ($mean(\alpha_2) = -1.122$). As for the *Biotic Coefficient*, a decrease in the first period is also found, but in a smaller scale ($mean(\alpha_2) = -0.767$) than the ones obtained for the remaining variables.

Numerical Summaries - Summer and Autumn

Considering the models for Summer and Autumn, there is not much differences from the model for Winter. Tables with all the posterior statistics obtained for each parameter considered for all the fitted models can be found in Appendix D.

4.4 Inter-individual differences

Besides the analysis of the variables as a whole, these models allow individual analysis of each one of the experimental units (sampling stations). For this purpose, each one of the 24 sampling stations have an initial level and slope estimated, represented by 48 extra parameters, besides the original ones described previously (see Table 4.4). As expected, the estimated trajectories for each sampling station have a similar shape as the mean trajectory, since the shape parameters are unique. The high variability observed in *Total Abundance* and *Taxonomic Richness* with the increase of organisms and taxa is due to the fact that these two variables were estimated using the negative binomial distribution.

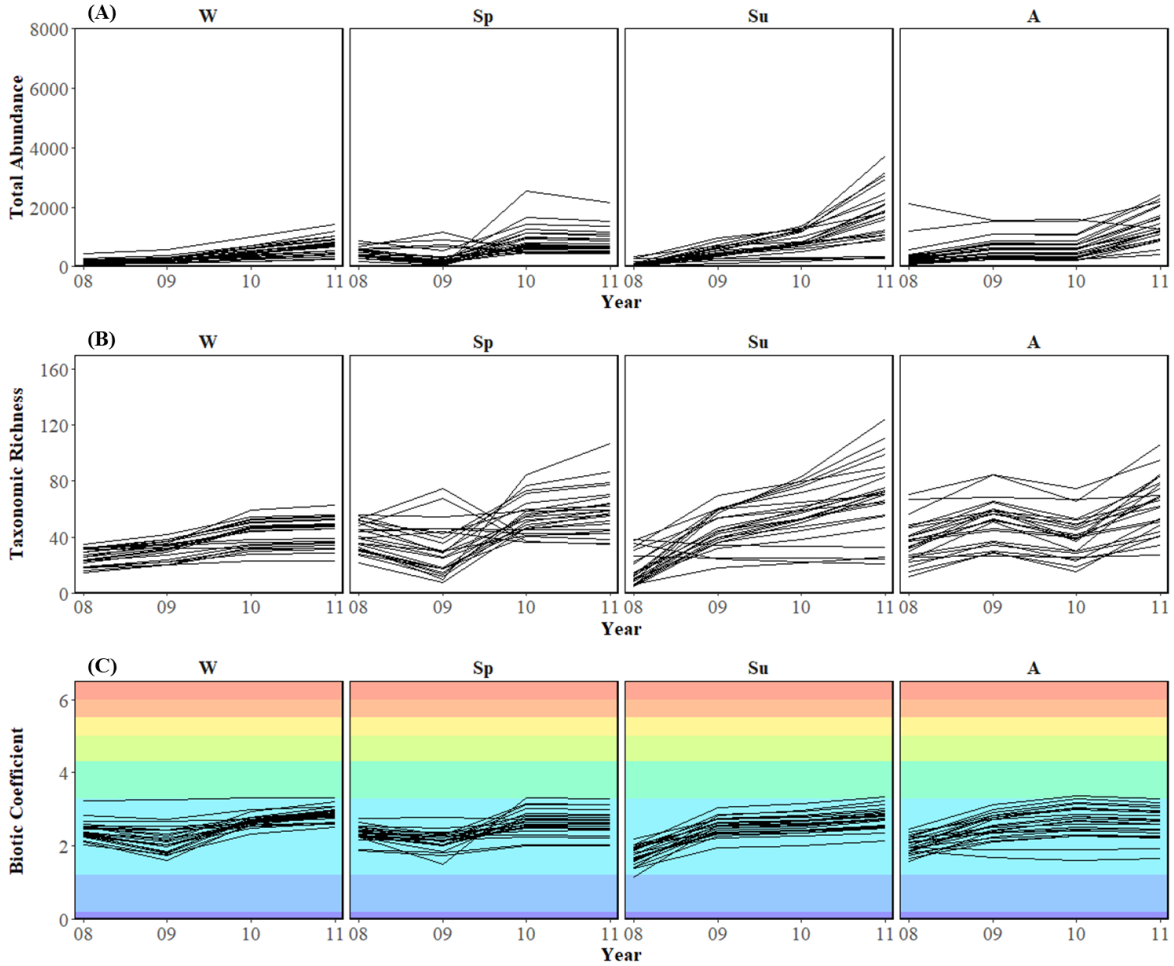

Figure 4.4: Person-specific estimates obtained for each one of the sampling stations. *Total Abundance* (A); *Taxonomic Richness* (B); and *Biotic Coefficient* (C).

4.5 Sensitivity Analysis

A sensitivity analysis was performed considering the same models and parameters described above, changing the prior distribution chosen for the shape parameters, since these are the main object of study when using the LGCMs. In this case, the Uniform prior distribution with a high variance (U[-1000, 1000]) was selected. The results obtained were the same as those presented above, showing that the selection of the Normal distribution for α_t was not a limiting option for the data.

5 | Final Remarks

In this final chapter, the main issues concerning the application of LGCMs to the study of ecological monitoring data will be addressed, as well as a simple discussion of the results obtained for the biological communities involved.

5.1 Discussion

Long-term and spatio-temporal changes in biological communities are topics that have been more and more investigated over the last years [15, 62, 63]. The impact of human activities in benthic communities has also been a concern in some countries where pollution levels, fishing rates and overexploitation of resources raised over time [64, 65]. Although several of these studies have used simple models to try to describe the variations found, many rely on more cursory statistical analysis in order to test the significance of differences over time and space. These are the reasons why this work presents an innovative approach to the long-term study of biological communities.

The results presented here not only show the temporal distribution of the benthic communities on these two sites, but also provide a description of the LGCMs that can be used to obtain, describe and quantify that temporal distribution. The results obtained by carrying out this work can now answer the questions initially asked. Regarding the comparison of the benthic communities of the two sites, no significant differences were found in the total abundance of organisms, the number of taxa found or in terms of ecological water quality. This lack of significance shows that, although the construction of a WWTP has interfered with the environment, it does not seem to have had a negative effect on the surrounding communities.

Although the analysis of these results leads to the conclusion that there is no evidence of an influence of WWTP activity on biological communities, no decisions can be made based on these results since the structure and composition of these communities can change very easily. In order to understand the natural fluctuation pattern of this structure, longer series are necessary so that the effect of natural change in the communities is captured and an erroneous attribution of responsibility is not made to external interventions [66].

Benthic macro invertebrates are organisms that react quickly to negative inputs into the system by disappearing from places where characteristics are unfavourable. However, for these communities to recover from abrupt events, the phenomenon of ecological succession must be completed. In drastic cases where all organisms disappear from a site due to external pressures (excessive pollution, overexploitation of resources, contamination by chemical agents, etc.), thus obtaining an azoic classification, recovery after the removal of that pressure can take several years. The process of ecological succession in this case involves the appearance of species that are tolerant to this pressure, followed by the slow resurgence of species that are less tolerant until the community is made up of organisms from different taxonomic groups

and with different levels of tolerance. When one is in the presence of a macro invertebrate community with these characteristics, one can say that the site has fully recovered from the destructive past event [67].

5.2 Conclusion

Regarding the expectations at the beginning of this work, it can be said that none of the hypotheses put forward were true. Looking at the variation in the total abundance of organisms over time, a considerable increase was observed in the last periods considered. This might mean that the communities under study are in a state of growth and not of stabilization. Although this can be stated by looking at the graphs obtained either through models' results or the actual data, there is no clear statistical quantification. This may be due to the use of data from only eight years of monitoring, *i.e.* if the time period is extended, the hypothesised trends may or may not be observed, but a more robust conclusion could be taken. The same can be said for the amount of taxa found in these sampled stations. Although a slight increase could be observed at the end of the time period considered, this growth is not statistically significant. There is, however, a need to consider something in relation to the results obtained in this work. Although the hypotheses initially put forward may not have been verified as a result of the short period of time used, the decrease in organisms identified in 2008 may also have considerably affected the communities, which did not had time to recover. A final comment regarding this abrupt decrease of organisms in the estuary is needed. Although one might think that could result from organic matter discharges in high quantities by the WWTP, it seems to not be the case as there is no increase in the total amount of organic matter in the sediments surrounding this infrastructure and also a break in the number and diversity of organisms has been observed in *Porto do Buxo* (where there are no more discharges to the aquatic environment).

In terms of quantification of the sites' ecological quality, a stabilisation in both of them was observed as the slopes obtained for each site separately and together were very close to 0. What can be expected from these communities, through the analysis of these results, will be a continuous evolution until further stabilization if there are no more abrupt events that will cause a decrease in the number of organisms.

The use of LGCMs has proved to be very useful in describing and capturing the shape and trajectory of the main variables presented in this study. Their use in ecological monitoring data seems then to allow a description, prediction and quantification of changes, patterns and possible events of disruption of biological communities. A special note about the effect of seasonality or any other patterns in the data is needed. Regarding these aspects, it was concluded that, separating the data with respect to the season (minor dependency identified in the exploratory analysis of the data) enabled the analysis of the data through the LGCMs. Other considerations regarding these models must be made. Since the main trends identified by the models were linear, and considering the extreme values observed in the data, there might be some non-linear latent effects which were not captured by the fitted models. A possibility might be to consider smoothing functions (*e.g.* splines) in order to capture such fluctuations.

5.3 Future Work

The use of Latent Growth Curve Models can have two purposes: the description of unknown or limited longitudinal data, or the incorporation of trends found in complex biological systems. In this work, only the first one was achieved. In the future, one of the objectives is to develop a biological systems' simulator where trajectories described using these methods will be used, but also incorporating other environmental variables that may prove relevant. The use of a larger time window is also one of the

options for the extension of this work. For the implementation of these methods on a larger time scale, more sampling stations would be needed. With the development of new projects in this scientific area, the modelling difficulties should be taken into account when performing the experimental design.

These models have shown to be helpful in describing and interpreting data from ecological quality assessment, and there are other areas where they may be useful. Some examples include: modelling and predicting the changes on the phytoplankton communities in estuaries and coastal areas, determining patterns in phytoplankton communities, monitoring the effect of intensive bivalve aquaculture, assessing the impact of new constructions near marine areas and many others.

Bibliography

[1] Cropper M. and Griffiths C. (1994), The Interaction of Population Growth and Environmental Quality, *The American Economic Review*, Vol. **84**, no. 2, pp. 250-254

[2] Harte J. (2007), Human population as a dynamic factor in environmental degradation, *Population and Environment*, Vol. **28**, no. 4-5, pp. 223-236

[3] Borja Á., Dauer D., Elliott M. and Simenstad C. A. (2010), Medium- and Long-term Recovery of Estuarine and Coastal Ecosystems: Patterns, Rates and Restoration Effectiveness, *Estuaries and Coasts*, Vol. **33**, Issue 6, pp. 1249-1260

[4] Arrow K., Bolin B., Costanza R., Dasgupta P., Folke C., Holling C.S., Jansson B., Levin S., Mäler K., Perrings C. and Pimentel D. (1995), Economic growth, carrying capacity, and the environment, *Ecological Economics*, Vol. **15**, no. 2, pp. 91-95

[5] Shafik N. (1994), Economic development and Environmental quality: an econometric analysis, *Oxford Economic Papers*, Vol. **46**, pp. 757-773

[6] World Bank, Economical indicators, Accessed 12th Jul. 2019, http://data.worldbank.org/country/portugal?view=chart

[7] Carson R. (1962), Silent Spring, *Crest Book*

[8] Union of Concerned Scientists (1997), World Scientists' Warning for Humanity, *Cambridge*

[9] European Commission (2010), EUROPE 2020: A strategy for smart, sustainable and inclusive growth

[10] Smith J. G., Brandt C. C. and Christensen S. W. (2011), Long-Term Benthic Macroinvertebrate Community Monitoring to Assess Pollution Abatement Effectiveness, *Environmental Management*, Vol. **47**, pp. 1077-1095

[11] Barbour M. T., Gerritsen J., Snyder B. D. and Stribling J. B. (1999), Rapid Bioassessment Protocols for Use in Streams and Wadeable Rivers: Periphyton, Benthic Macroinvertebrates and Fish, *EPA 841-B-99-002. U.S. Environmental Protection Agency; Office of Water*, Washington, D.C., Second Edition

[12] Hodkinson I. D. and Jackson J. K. (2005), Terrestrial and Aquatic Invertebrates as Bioindicators for Environmental Monitoring, with Particular Reference to Mountain Ecosystems, *Environmental Management*, Vol. **35**, no. 5, pp. 649-666

[13] Dauer D. (1993), Biological criteria, environmental health and estuarine macrobenthic community structure, *Marine Pollution Bulletin*, Vol. **25**, no. 5, pp. 249-257

[14] Hauer F. R. and Lamberti G. A. (2017), Methods in Stream Ecology; Volume 1: Ecosystem Structure, *Academic press*, Third Edition

[15] Chainho P., Silva G.,Lane F., Costa J.L., Pereira T., Azeda C., Almeida P. R., Metelo I. and Costa M. J. (2010), Long-Term Trends in Intertidal and Subtidal Benthic Communities in Response to Water Quality Improvement Measures, *Estuaries and Coasts*, Vol. **33**, pp. 1314-1326

[16] Eleftheriou A. and McIntyre A. (2005), Methods for the study of marine benthos, *Blackwell Science Ltd*, Third Edition

[17] Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy

[18] Hering D., Borja Á., Carstensen J., Carvalho L., Elliott M., Feld C. K., Heiskanen A., Johnson R. K., Moe J., Pont D., Solheim A. L. and Bund W. (2010), The European Water Framework Directive at the age of 10: A critical review of the achievements with recommendations for the future, *Science of the Total Environment*, Vol. **408**, pp. 4007-4019

[19] Working Group 2.2 on Heavily Modified Water Bodies (2003), Identification and Designation of Heavily Modified and Artificial Water Bodies, *European Commission*, Guidance document No. 4

[20] Borja Á., Miles A., Occhipinti-Ambrogi A. and Berg T. (2009), Current status of macroinvertebrate methods used for assessing the quality of European marine waters: implementing the Water Framework Directive, *Hydrobiologia*, Vol. **633**, pp. 181-196

[21] Council Directive 91/271/EEC of 21 May 1991 concerning urban waste-water treatment

[22] Decree-law 152/97 of 19th June 1997- Republic Diary No. 139/1997

[23] Alden R.W., Weisberg S.B., Ranasinghe J.A. and Dauer D.M. (1997), Optimizing temporal sampling strategies for benthic environmental monitoring programs, *Marine Pollution Bulletin*, Vol. **34**, pp. 913-922

[24] Reiss H. and Krüncke I. (2005), Seasonal variability of benthic indices: an approach to test the applicability of different indices for ecosystem quality assessment, *Marine Pollution Bulletin*, Vol. **50**, pp 1490-1499

[25] Chainho P., Costa J.L., Chaves M.L., Dauer D.M. and Costa M.J. (2007), Influence of seasonal variability in benthic invertebrate community structure on the use of biotic indices to assess the ecological status of a Portuguese estuary, *Marine Pollution Bulletin*, Vol. **54**, pp. 1586-1597

[26] Borja Á., Franco J. and Peréz V. (2000), A Marine Biotic Index to Establish the Ecological Quality of Soft-Bottom Benthos Within European Estuarine and Coastal Environments, *Marine Pollution Bulletin*, Vol. **40**, no. 12, pp. 1100-1114

[27] Pearson, T. H. and Rosenberg, R. (1978), Macrobenthic succession in relation to organic enrichment and pollution of the marine environment, *Oceanography and Marine Biology*, Vol. **16**, pp. 229-311

[28] Vaughan H., Brydges T., Fenech A. and Lumb A. (2001), Monitoring long-term biological changes through the ecological monitoring and assessment network: Science-based and Policy relevant, *Environmental Monitoring and Assessment*, Vol. **67**, pp. 3-28

[29] Vesilind P. A., Peirce J. J. and Weiner R. F. (1990), Environmental Pollution and Control, chapter Wastewater Treatment *Butterworth-Heinemann*, Third Edition

[30] Artiola J.F., Pepper I. L. and Brusseau M.L. (2004), Environmental Monitoring and Characterization, chapter Chemical Contaminants, *Academic Press*

[31] Bilyard G. R. (1987), The Value of Benthic Infauna in Marine Pollution Monitoring Studies, *Marine Pollution Bulletin*, Vol. **18**, no. 11, pp. 581-585

[32] Costa J.L., Costa M.J., Cabral H., Almeida P.R., Caçador M.I., Silva G., Tavares M.J., Medeiros J.P., Sá E., Teixeira C. and Pedro S. (2017), Monitorização das comunidades marinhas e caracterização das atividades de pesca com Xávega na zona costeira do Concelho de Almada, Final Report, MARE - Marine and Environmental Science Center, Lisbon

[33] APA – Agência Portuguesa do Ambiente (2016), Plano de Gestão de Região Hidrográfica, Região Hidrográfica do Tejo e Ribeiras do Oeste (RH5). Parte 2 – Caracterização e Diagnóstico, Hydrographic Region Management Plan 2006/2021, Lisbon

[34] Rodrigues M., Rosa A., Cravo A, Fortunado A.B. and Jacob J. (2017), Report 1 - Characterization of the study areas: Tagus estuary and Ria Formosa, Laboratório Nacional de Engenharia Civil, Lisbon

[35] IPMA - Instituto Português do Mar e da Atmosfera, Accessed 9th Sep. 2019, http://www.ipma.pt/pt/oclima/monitorizacao/

[36] AZTI (2019), AMBI: AZTI's Marine Biotic Index [software], https://ambi.azti.es/descarga-de-ambi/

[37] Grall J. and Glèmarec M. (1997), Using Biotic Indices to Estimate Macrobenthic Community Perturbations in the Bay of Brest, *Estuarine, Coastal and Shelf Science*, Vol. **44**, pp. 43-45

[38] Hily C., Le Bis H. and Glémarec M. (1986), Impacts biologiques des emissaires urbains sur les ecosystemes benthiques, *Océanis*, Vol. **12**, no. 6, pp. 419-426

[39] Majeed S. A. (1987), Organic Matter and Biotic Indices on the Beaches of North Brittany, *Marine Pollution Bulletin*, Vol. **18**, no. 9, pp. 490-495

[40] Fournier J., Gallon R. K. and Paris R. (2014), G2Sd: a new R package for the statistical analysis of unconsolidated sediments, *Geomorphologie: relief, processus, environnement*, no. 1, pp. 73-78

[41] Google, (n.d.), [Google Maps distance measurements in the Tagus Estuary, Portugal], Accessed 12th Jun. 2019, https://www.google.pt/maps/@38.6765934,-9.2237552,15.75z?hl=pt-PT

[42] Grün B., Kosmidis I. and Zeileis A. (2012), Extended Beta Regression in R: Shaken, Stirred, Mixed, and Partitioned, *Journal of Statistical Software*, Vol. **48**, no. 11, pp. 1-25

[43] Amaral Turkman M. A. and Silva G. L. (2000), Modelos Lineares Generalizados - da teoria à prática, *VIII Congresso Nacional da Sociedade Portuguesa de Estatística*

[44] Faraway J. J. (2014), Linear Models with R, Chapman and Hall/CRC, Second Edition

[45] Paulino C., Turkman M. A., Murteira B. J. F. and Silva G. L. (2018), *Estatística Bayesiana*, *Fundação Calouste Gulbenkian*, Second Edition

[46] Bayes T. (1763), LII An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, *A. M. F. R. S, Philosophical Transactions Royal Society*, Vol. **53**, pp. 370 - 418

[47] Malécot, G. (1947), Annotated translation by D. Gianola of: Les criteres statistiques et la subjectivite de la connaisance scientifique (Statistical methods and the subjective basis of scientific knowledge), *Genetics, Selection, Evolution*, Vol. **31**, pp. 269-298

[48] Sorensen D. and Gianola D. (2002), Likelihood, Bayesian and MCMC methods in quantitative genetics, *New York: Springer-Verlag*, Sixth Edition

[49] Jeffreys, H. (1961), Theory of probability, *Oxford University Press*, New York, NY

[50] Shi D. and Tong X. (2017), The Impact of Prior Information on Bayesian Latent Basis Growth Model Estimation, *SAGE Open*, Vol. **7**, no. 3, pp. 1-14

[51] Metropolis N., Rosenbluth A. W., Rosenbluth M. N., Teller A. H. and Teller E. (1953), Equations of state calculations by fast computing machines, *Journal of Chemical Physics*, Vol. **21**, pp. 1087-1092

[52] Hastings W. K. (1970), Monte Carlo sampling methods using Markov chains and their application, *Biometrika*, Vol. **57**, pp. 97-109

[53] Geman, S. and Geman D. (1984), Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. **6**, pp. 721-741

[54] Resnick, S. (2002), Adventures in Stochastic Processes, *Springer Science+Business Media, LLC*

[55] Zhang Z., Hamagami F., Wang L., Nesselroade J. R. and Grimm K. J. (2007), Bayesian analysis of longitudinal data using growth curve models, *International Journal of Behavioural Development*, Vol. **31**, no. 4, pp. 374-383

[56] Wanga Y. and Daniels M. J. (2013), Bayesian modeling of the dependence in longitudinal data via partial autocorrelations and marginal variances, *Journal of Multivariate Analysis*, Vol. **116**, pp. 130-140

[57] McArdle J. J. and Grimm K. J. (2010), Longitudinal Research with Latent Variables, chapter Five Steps in Latent Curve and Latent Change Score Modeling with Longitudinal Data, *Springer-Verlag Berlin Heidelberg*

[58] Meredith, W. and Tisak, J. (1990), Latent curve analysis, *Psychometrika*, Vol. **55**, pp. 107-122

[59] Nesselroade J. R. and Baltes P. B. (1979), Longitudinal Research in the study of Behaviour and Development, chapter History and rationale of longitudinal research, *New York: Academic Press*

[60] Oravecz Z. and Muth C. (2018), Fitting growth curve models in the Bayesian framework, *Psychonomic Bulletin Review*, Vol. **25**, pp. 235-255

[61] Gelman A., Carlin J. B., Stern H. S., Dunson D. B., Vehtari A. and Rubin D. B. (2014), Bayesian Data Analysis, *Taylor & Francis Group, LLC*, Third Edition

[62] Nalepa T. F., Fanslow D. L., Pothoven S. A., Foley III A. J. and Lang G. A. (2007), Long-term Trends in Benthic Macroinvertebrate Populations in Lake Huron over the Past Four Decades, *Journal of Great Lakes Research*, Vol. **33**, no. 2, pp. 421-436

[63] Tsikopoulou I., Moraitis M. L., Geropoulos A., Papadopoulou K. N., Papageorgiou N., Plaiti W., Smith C. J., Karakassis I. and Eleftheriou A. (2018), Long-term changes in the structure of benthic communities: revisiting a sampling transect in Crete after 24 years, *Marine Environmental Research*, Vol. **144**, pp. 9-19

[64] Portugal A. B., Carvalho F. L., Carneiro P. B. M. and Soares M. O. (2016), Increased anthropogenic pressure decreases species richness in tropical intertidal reefs, *Marine Environmental Research*, Vol. **120**, pp. 44-54

[65] Martinelli J. C., Soto L. P., González J. and Rivadeneira M. M. (2017), Benthic communities under anthropogenic pressure show resilience across Quaternary, *Royal Society Open Science*, Vol. **4**, pp. 1-12

[66] Gray J. S. and Christie H. (1983), Predicting long-term changes in marine benthic communities, *Marine Ecology - Progress Series*, Vol. **13**, pp. 87-94

[67] Chang C. C. and Turner B. L. (2019), Ecological sucession in a changing world, *Journal of Ecology*, Vol. **107**, pp. 503-509

A | Annual and Seasonal Influence tests' results

Table A.1: Values obtained for the coefficient estimative (top value) of all parameters involved and the test statistics (bottom left value) and respective p-value (bottom right value in parentheses) of Wald's test for the three models constructed (one for each time interval considered) for each one of the response variables and sites (N - *Total Abundance*; S - *Taxonomic Richness*; BC - *Biotic Coefficient*).

		Porto do Buxo		
		N	**S**	**BC**
(Intercept)		6.129 38.370 ($\ll$ 0.05)	3.459 38.486 ($\ll$ 0.05)	3.141 23.697 ($\ll$ 0.05)
Year	2005	-0.121 -0.540 (> 0.0167)	0.400 3.194 (< 0.0167)	-0.837 -4.494 ($\ll$ 0.0167)
	2006	0.254 1.131 (> 0.0167)	0.583 4.672 ($\ll$ 0.0167)	-0.707 -3.798 ($\ll$ 0.0167)
	2007	0.357 1.593 (> 0.0167)	0.684 5.494($\ll$ 0.0167)	-1.182 -6.350 ($\ll$ 0.0167)
(Intercept)		6.487 38.713 ($\ll$ 0.05)	4.144 41.909 ($\ll$ 0.05)	1.959 15.005 ($\ll$ 0.05)
Year	2008	-0.922 -3.887 ($\ll$ 0.025)	-0.732 -5.174 ($\ll$ 0.025)	0.097 0.523 (< 0.025)
(Intercept)		5.565 34.713 ($\ll$ 0.01)	3.411 36.796 ($\ll$ 0.01)	2.056 20.131 ($\ll$ 0.05)
Year	2009	0.675 2.957 (< 0.0167)	0.417 3.212 (< 0.0167)	0.431 2.985 (< 0.0167)
	2010	0.855 3.748 ($\ll$ 0.0167)	0.410 3.151 ($\ll$ 0.0167)	0.910 6.304 ($\ll$ 0.0167)
	2011	1.256 5.259 ($\ll$ 0.0167)	0.517 3.806 ($\ll$ 0.0167)	0.803 5.303 ($\ll$ 0.0167)
		Portinho da Costa		
		N	**S**	**BC**
(Intercept)		5.251 31.849 ($\ll$ 0.05)	3.123 31.525 ($\ll$ 0.05)	2.139 26.526 ($\ll$ 0.05)
Year	2005	0.719 3.109 (< 0.0167)	0.407 2.946 (< 0.0167)	-0.064 -0.565 (> 0.0167)
	2006	0.900 3.897 ($\ll$ 0.0167)	0.557 4.040 ($\ll$ 0.0167)	0.007 0.059 (> 0.0167)
	2007	1.221 5.241 ($\ll$ 0.0167)	0.761 5.490 ($\ll$ 0.0167)	-0.430 -3.771 ($\ll$ 0.0167)
(Intercept)		6.472 38.294 ($\ll$ 0.05)	3.884 41.282 ($\ll$ 0.05)	1.709 24.623 ($\ll$ 0.05)
Year	2008	-0.625 -2.626 (< 0.025)	-0.516 -3.872 ($\ll$ 0.025)	0.402 4.121 ($\ll$ 0.025)
(Intercept)		5.847 42.612 ($\ll$ 0.01)	3.367 43.723 ($\ll$ 0.05)	2.112 29.079 ($\ll$ 0.05)
Year	2009	0.004 0.020 (> 0.0167)	0.234 2.165 (> 0.00167)	0.147 1.438 (> 0.0167)
	2010	0.751 3.886 (< 0.0167)	0.524 4.882 ($\ll$ 0.0167)	0.406 3.973 ($\ll$ 0.0167)
	2011	1.255 5.901 ($\ll$ 0.0167)	0.775 5.152 (< 0.0167)	0.581 5.160 ($\ll$ 0.0167)

Table A.2: Values obtained for the coefficient estimative (top value) of all parameters involved and the test statistics (bottom left value) and respective p-value (bottom right value in parentheses) of Wald's test for each one of the response variables and sites (MGS - *Mean Grain Size*; TOM - *Total Organic Matter*).

		Porto do Buxo		*Portinho da Costa*	
		MGS	**TOM**	**MGS**	**TOM**
(Intercept)		669.16 10.481 ($\ll$ 0.05)	-3.709 -54.630 ($\ll$ 0.05)	461.31 9.527 ($\ll$ 0.05)	-3.660 59.227 ($\ll$ 0.05)
Year	2005	25.43 0.282 (> 0.007)	0.142 1.530 (> 0.007)	57.42 0.864 (> 0.007)	0.188 2.270 (> 0.007)
	2006	34.49 0.382 (> 0.007)	0.289 3.209 (< 0.007)	61.61 0.907 (> 0.007)	0.288 3.544 (< 0.007)
	2007	-83.00 -0.913 (> 0.007)	-0.066 -0.681 (> 0.007)	21.65 0.316 (> 0.007)	-0.106 -1.190 (> 0.007)
	2008	110.55 1.224 (> 0.007)	0.054 0.574 (> 0.007)	62.92 0.923 (> 0.007)	0.269 3.275 (< 0.007)
	2009	248.80 2.755 (< 0.007)	0.319 3.564 ($\ll$ 0.007)	112.77 1.661 (> 0.007)	0.187 2.257 (> 0.007)
	2010	351.24 3.890 ($\ll$ 0.007)	0.277 3.066 (< 0.007)	470.23 6.925 ($\ll$ 0.007)	0.363 4.514 ($\ll$ 0.007)
	2011	198.91 2.203 (> 0.007)	0.202 2.202 (> 0.007)	305.43 4.498 ($\ll$ 0.007)	0.432 5.448 ($\ll$ 0.007)

Table A.3: Values obtained for the coefficient estimative (top value) of all parameters involved and the test statistics (bottom left value) and respective p-value (bottom right value in parentheses) of Wald's test for each one of the response variables and sites (N - *Total Abundance*; S - *Taxonomic Richness*), BC - *Biotic Coefficient*.

		Porto do Buxo		
		N	**S**	**BC**
(Intercept)		5.999 51.542 ($\ll$ 0.05)	3.605 53.286 ($\ll$ 0.05)	2.785 30.071 ($\ll$ 0.05)
Season	Sp	0.302 1.830 (> 0.0167)	0.264 2.758 (< 0.0167)	-0.289 -2.200 (< 0.0167)
	Su	0.310 1.863 (> 0.0167)	0.006 0.440 (> 0.0167)	-0.528 -3.984 ($\ll$ 0.0167)
	A	0.536 3.224 (< 0.0167)	0.355 3.695 ($\ll$ 0.0167)	-0.270 -2.038 (> 0.0167)
		Portinho da Costa		
		N	**S**	**BC**
(Intercept)		5.950 51.964 ($\ll$ 0.05)	3.556 52.930 ($\ll$ 0.05)	2.365 40.316 ($\ll$ 0.05)
Season	Sp	0.409 2.546 (< 0.0167)	0.172 1.834 (> 0.0167)	-0.154 -1.869 (> 0.0167)
	Su	0.358 2.181 (< 0.01675)	0.179 1.863 (> 0.0167)	-0.320 -3.803 ($\ll$ 0.0167)
	A	0.379 2.316 (< 0.0167)	0.121 1.257 (> 0.0167)	-0.239 -2.848 (< 0.0167)

Table A.4: Values obtained for the coefficient estimative (top value) of all parameters involved and the test statistics (bottom left value) and respective p-value (bottom right value in parentheses) of Wald's test for each one of the response variables and sites (MGS - *Mean Grain Size*; TOM - *Total Organic Matter*).

		Porto do Buxo		*Portinho da Costa*	
		MGS	**TOM**	**MGS**	*TOM*
(Intercept)		769.47 14.243 ($\ll$ 0.01)	-3.593 -74.243 ($\ll$ 0.01)	596.421 16.130 ($\ll$ 0.01)	-3.468 -80.198 ($\ll$ 0.01)
Season	Sp	51.31 0.763 (> 0.1)	0.076 1.133 (> 0.1)	-5.814 -0.112 (> 0.1)	0.029 0.462 (> 0.1)
	Su	-51.42 -0.768 (> 0.1)	0.056 0.832 (> 0.1)	-12.251 -0.236 (> 0.1)	0.041 0.688 (> 0.1)
	A	45.33 0.677 (> 0.1)	0.047 0.704 (> 0.1)	28.294 0.546 (> 0.1)	0.029 0.488 (> 0.1)

B | Data Description

B.1 *Porto do Buxo*

Table B.1: Descriptive statistics obtained for the three biological variables (N - *Total Abundance*; S - *Taxonomic Richness*; BC - *Biotic Coefficient*) in *Porto do Buxo* (mean and standard deviation (SD)). Dashed lines indicate the year division in which the four sampling seasons are indicated (W - Winter, Sp - Spring, Su - Summer, A - Autumn). Incomplete data points are indicated with parentheses.

Season	n	Mean (N)	SD (N)	Mean (S)	SD (S)	Mean (BC)	SD (BC)
W	9	642.111	491.250	36.556	10.212	3.453	0.324
Sp	(8)	96.500	115.321	17.125	13.757	2.418	1.147
Su	9	361.111	305.671	36.333	17.220	2.785	0.843
A	9	696.222	585.968	35.556	11.717	3.827	0.841
W	9	238.667	378.672	26.333	18.166	2.603	1.159
Sp	9	509.889	421.451	50.667	20.809	2.338	0.409
Su	9	562.000	281.750	64.778	19.665	2.179	0.394
A	9	316.222	258.475	47.889	18.711	2.097	0.561
W	9	245.667	244.628	35.444	13.584	2.535	0.480
Sp	9	853.889	578.135	66.556	14.621	2.723	0.395
Su	9	895.889	506.301	80.222	17.484	2.479	0.349
A	9	370.889	327.989	45.556	18.514	1.999	0.733
W	9	584.444	261.910	54.667	13.019	2.796	0.449
Sp	9	805.556	325.401	79.111	14.348	2.167	0.438
Su	9	355.778	342.890	42.000	32.955	0.984	0.542
A	9	878.778	500.648	76.333	19.352	1.888	0.769
W	9	155.778	115.531	27.889	11.837	2.455	0.346
Sp	9	374.444	161.550	40.333	8.803	2.442	0.576
Su	9	85.111	128.195	15.333	16.409	1.443	0.873
A	9	428.889	834.208	37.667	21.909	1.882	0.402
W	9	283.444	206.540	31.556	13.538	2.519	1.015
Sp	9	529.667	764.040	37.778	30.091	2.025	0.706
Su	9	551.889	490.694	51.667	24.068	2.567	0.451
A	9	685.778	237.506	63.000	10.037	2.835	0.373
W	9	422.556	323.306	41.556	17.889	2.914	0.730
Sp	9	605.889	406.214	45.444	16.241	3.064	0.480
Su	9	593.444	390.067	46.889	20.521	2.893	0.605
A	9	834.333	562.329	48.667	23.696	2.993	0.316
W	9	649.778	401.198	40.333	10.025	3.005	0.367
Sp	9	533.667	315.392	42.667	15.700	2.783	0.351
Su	(6)	1207.167	935.516	58.167	22.104	2.966	0.229
A	(6)	1601.833	943.961	71.500	34.634	2.646	0.618

B.2 *Portinho da Costa*

Table B.2: Descriptive statistics obtained for the three biological variables in *Portinho da Costa* (mean and standard deviation (SD)). Dashed lines indicate the year division in which the four sampling seasons are indicated (W - Winter, Sp - Spring, Su - Summer, A - Autumn). Incomplete data points are indicated with parentheses.

Season	n	Mean (N)	SD (N)	Mean (S)	SD (S)	Mean (BC)	SD (BC)
W	(13)	331.000	375.765	30.923	23.243	2.424	0.662
Sp	15	70.800	139.645	11.400	11.287	1.839	1.078
Su	15	129.000	152.819	25.467	17.150	2.050	0.518
A	15	250.667	334.076	24.133	15.688	2.281	0.486
W	15	247.000	236.296	26.333	13.399	2.658	0.560
Sp	15	325.333	525.715	28.867	21.267	1.990	0.420
Su	15	500.267	544.182	41.733	30.990	1.584	0.480
A	15	492.133	668.411	39.467	27.874	2.068	0.529
W	15	458.200	709.574	33.800	28.919	1.966	0.534
Sp	15	628.067	619.042	45.067	24.426	2.423	0.513
Su	15	674.467	552.044	53.467	29.845	2.304	0.297
A	15	116.133	100.323	26.133	14.995	1.891	0.454
W	(13)	487.077	562.079	46.000	28.917	2.109	0.509
Sp	15	593.000	684.776	52.533	34.692	2.025	0.327
Su	15	459.600	705.275	34.200	29.771	1.269	0.431
A	15	1025.600	1027.460	61.333	33.909	1.487	0.535
W	15	200.733	267.355	23.867	9.598	2.222	0.488
Sp	15	406.800	432.298	37.667	27.794	2.229	0.402
Su	(14)	105.429	150.447	19.143	15.150	1.904	0.487
A	15	655.400	1635.951	34.667	24.657	2.079	0.535
W	15	179.200	114.289	31.333	12.246	2.037	0.605
Sp	15	265.467	323.695	29.467	21.263	2.214	0.521
Su	15	494.933	454.812	39.400	19.353	2.518	0.853
A	15	450.200	381.277	46.333	23.868	2.267	0.640
W	15	459.733	299.653	44.200	14.804	2.620	0.367
Sp	15	1094.200	1880.135	54.333	29.509	2.516	0.468
Su	15	841.267	643.962	59.933	25.178	2.421	0.581
A	15	537.200	629.647	37.467	23.148	2.516	0.762
W	15	713.667	578.379	44.667	21.652	2.860	0.277
Sp	15	1236.600	1143.295	73.600	29.899	2.456	0.539
Su	(6)	2068.167	1006.556	87.833	31.276	2.693	0.454
A	(6)	1554.167	1066.141	57.167	23.259	2.869	0.546

C | LGCM's implemented in JAGS

The code to fit the models used in this thesis was adapted from Zhang *et. al* (2007) [55].

C.1 *Total Abundance* and *Taxonomic Richness*

```
model{
  # loop over individuals
  for(i in 1:N){
    # loop over measurement occasions
    for(t in 1:nrT){
    # Specifying the likelihood
    Y[i, t] ~ dnegbin(p[i, t], r)

    # probability of success inherent to the Negative Binomial
    p[i, t] <- r/(r + lambda[i, t])

    # Expected value of Y
    log(MuY[i, t]) <- LS[i, 1] + LS[i, 2] * alpha[t]

    lambda[i, t] <- MuY[i, t]
  }

  LS[i, 1:2] ~ dmnorm(Mu[i, 1:2], PrecisionMatrix[1:2, 1:2])

  for(i in 1:1){
    Mu[i, 1] <- MuL
    Mu[i, 2] <- MuS
  }

  # Dispersion parameter
  r ~ dunif(0, 50)

  # Shape parameter
  alpha[1] <- 0
  alpha[4] <- 1

  for(t in 2:3){
    alpha[t] ~ dnorm(0, 0.001)
  }

  # Specify prior distributions
  MuL ~ dnorm(0, 0.001)
  MuS ~ dnorm(0, 0.001)
```

```
  PrecisionMatrix[1:2, 1:2] ~ dwish(R[1:2, 1:2], 2)
  R[1, 1] <- 1
  R[2, 2] <- 1
  R[2, 1] <- R[1, 2]
  R[1, 2] <- 0

  CovMatrix[1:2, 1:2]  <- inverse(PrecisionMatrix[1:2, 1:2])
}
```

C.2 *Biotic Coefficient*

```
model{
  # loop over sampling units
  for(i in 1:N){
    # loop over measurement occasions
    for(t in 1:nrT){
      # Specifying the likelihood
      Y[i, t] ~ dnorm(MuY[i, t], tau_e)

      # Expected value of Y
      MuY[i, t] <- LS[i, 1] + LS[i, 2] * alpha[t]
    }

    LS[i, 1:2] ~ dmnorm(Mu[i, 1:2], PrecisionMatrix[1:2, 1:2])
    Mu[i, 1] <- MuL
    Mu[i, 2] <- MuS
  }

  # Dispersion parameter
  tau_e ~ dgamma(0.01, 0.01)
  sig2_e <- 1/tau_e

  # Shape parameter
  alpha[1] <- 0
  alpha[4] <- 1

  for(t in 2:3){
    alpha[t] ~ dnorm(0, 0.001)
  }

  # Specify prior distributions
  MuL ~ dnorm(0, 0.001)
  MuS ~ dnorm(0, 0.001)

  PrecisionMatrix[1:2, 1:2] ~ dwish(R[1:2, 1:2], 2)
  R[1, 1] <- 1
  R[2, 2] <- 1
  R[2, 1] <- R[1, 2]
  R[1, 2] <- 0

  CovMatrix[1:2, 1:2] <- inverse(PrecisionMatrix[1:2, 1:2])
}
```

D | Convergence Analysis

D.1 *Total Abundance*

Winter

Table D.1: Posterior statistics of the parameters of the models considered for the *Total Abundance*'s model over all sampling stations (sd - Posterior Standard Deviation, ETP - Equal Tail Probability, ESS - Effective Sample Size).

	mean	**sd**	**ETP (2.5%)**	**ETP (97.5%)**	**ESS**	$\hat{R}$
σ_L^2	0.384	0.173	0.114	0.731	9704	1.000
σ_{LS}	-0.106	0.151	-0.434	0.147	9388	1.000
σ_S^2	0.364	0.212	0.075	0.783	9377	1.000
μ_L	5.047	0.186	4.692	5.426	8908	0.999
μ_S	1.318	0.224	0.881	1.756	8671	0.999
α_2	0.202	0.132	-0.071	0.449	8759	1.000
α_3	0.712	0.113	0.496	0.940	9935	0.999
r	2.946	0.512	2.014	3.959	10000	1.000

Spring

Table D.2: Posterior statistics of the parameters of the models considered for the *Total Abundance*'s model over all sampling stations (sd - Posterior Standard Deviation, ETP - Equal Tail Probability, ESS - Effective Sample Size).

	mean	**sd**	**ETP (2.5%)**	**ETP (97.5%)**	**ESS**	$\hat{R}$
σ_L^2	0.338	0.168	0.094	0.680	8771	1.000
σ_{LS}	-0.167	0.169	-0.538	0.083	6203	0.999
σ_S^2	0.454	0.281	0.077	0.999	5659	0.999
μ_L	6.057	0.184	5.698	6.425	7057	1.000
μ_S	0.579	0.229	0.178	1.054	5176	1.000
α_2	-1.407	0.733	-2.837	-0.149	3651	1.000
α_3	1.142	0.354	0.506	1.855	7039	0.999
r	1.593	0.272	1.070	2.113	9693	1.000

Summer

Table D.3: Posterior statistics of the parameters of the models considered for the *Total Abundance*'s model over all sampling stations (sd - Posterior Standard Deviation, ETP - Equal Tail Probability, ESS - Effective Sample Size).

	mean	**sd**	**ETP (2.5%)**	**ETP (97.5%)**	**ESS**	$\hat{R}$
σ_L^2	1.968	0.927	0.412	3.855	7797	0.999
σ_{LS}	-2.526	1.292	-5.027	-0.259	7966	1.000
σ_S^2	4.175	2.126	0.606	8.371	7764	1.000
μ_L	3.806	0.365	3.100	4.512	8749	1.000
μ_S	3.390	0.525	2.414	4.470	8578	1.001
α_2	0.676	0.067	0.543	0.806	7176	0.999
α_3	0.820	0.072	0.679	0.964	6907	1.000
r	1.977	0.488	1.089	2.950	7740	1.000

Autumn

Table D.4: Posterior statistics of the parameters of the models considered for the *Total Abundance*'s model over all sampling stations (sd - Posterior Standard Deviation, ETP - Equal Tail Probability, ESS - Effective Sample Size).

	mean	**sd**	**ETP (2.5%)**	**ETP (97.5%)**	**ESS**	$\hat{R}$
σ_L^2	1.043	0.505	0.255	2.017	7846	1.000
σ_{LS}	-0.832	0.634	-2.126	0.119	6327	0.999
σ_S^2	1.331	1.064	0.092	3.398	6656	0.999
μ_L	5.596	0.294	5.003	6.155	7686	0.999
μ_S	1.577	0.417	0.772	2.411	68883	0.999
α_2	0.472	0.151	0.172	0.769	7650	0.999
α_3	0.455	0.162	0.125	0.760	7472	1.000
r	1.603	0.309	1.051	2.226	8478	0.999

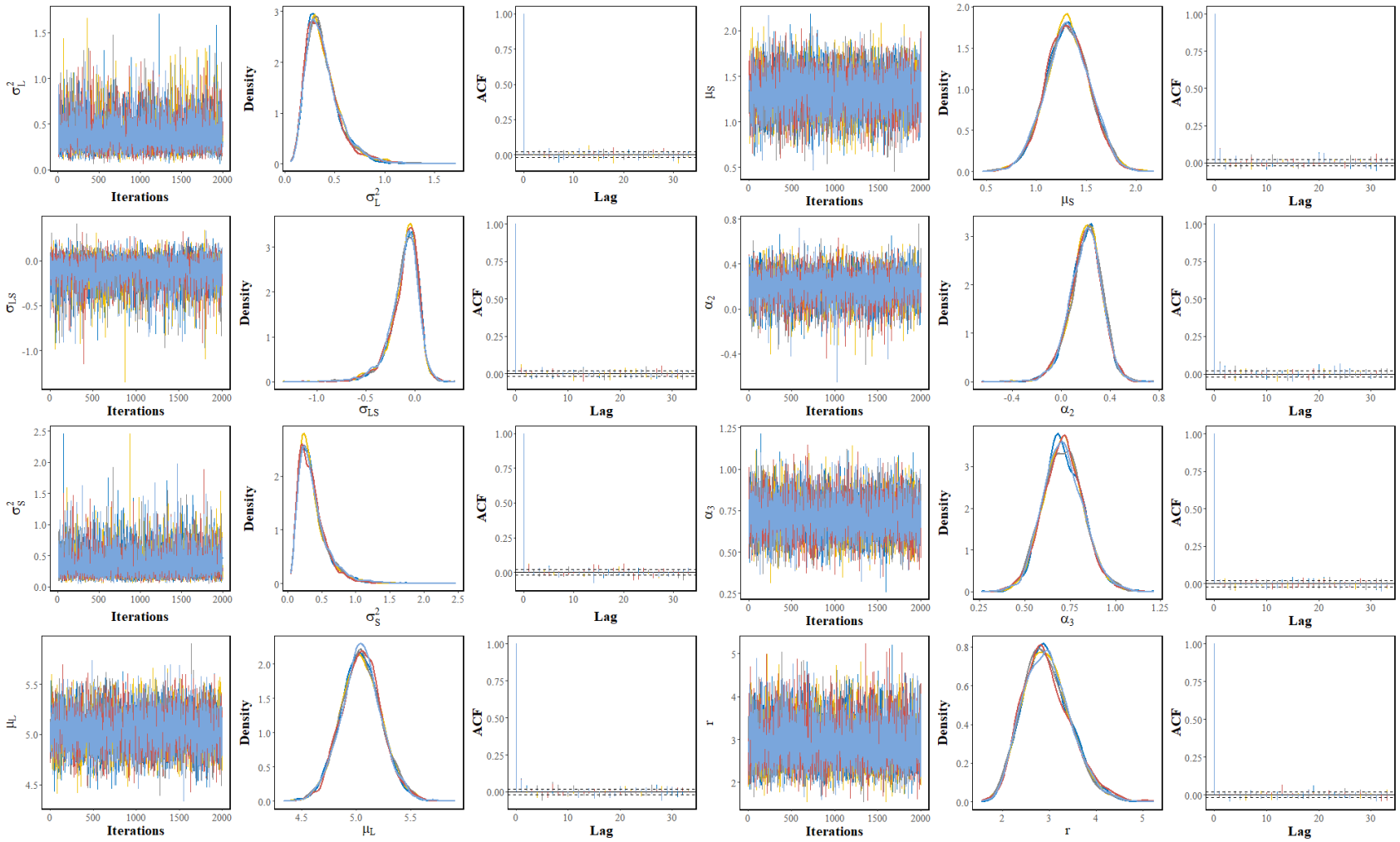

Figure D.1: Trace plots, density plots and ACF for the parameters involved on the modelling of *Total Abundance*.

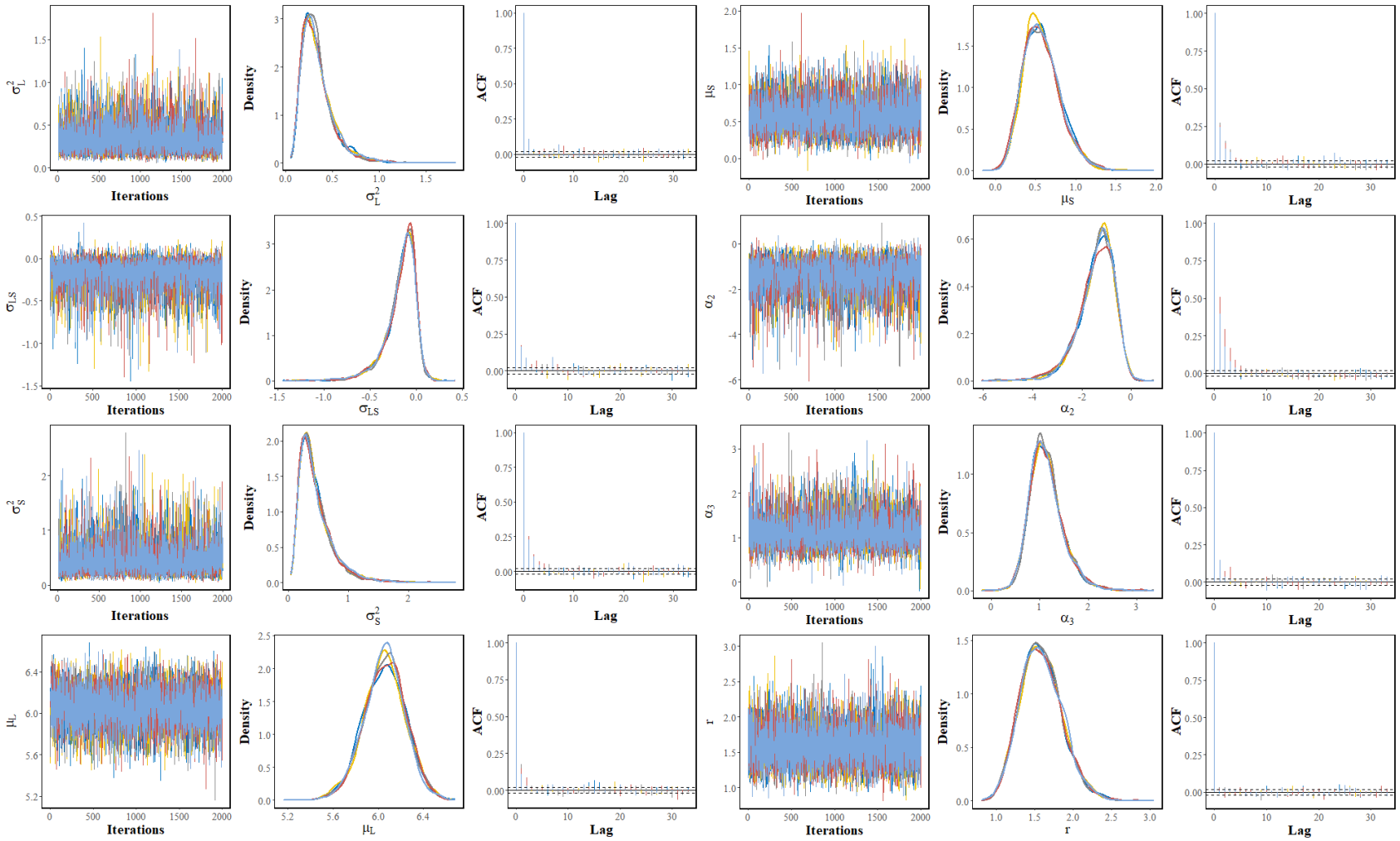

Figure D.2: Trace plots, density plots and ACF for the parameters involved on the modelling of *Total Abundance*.

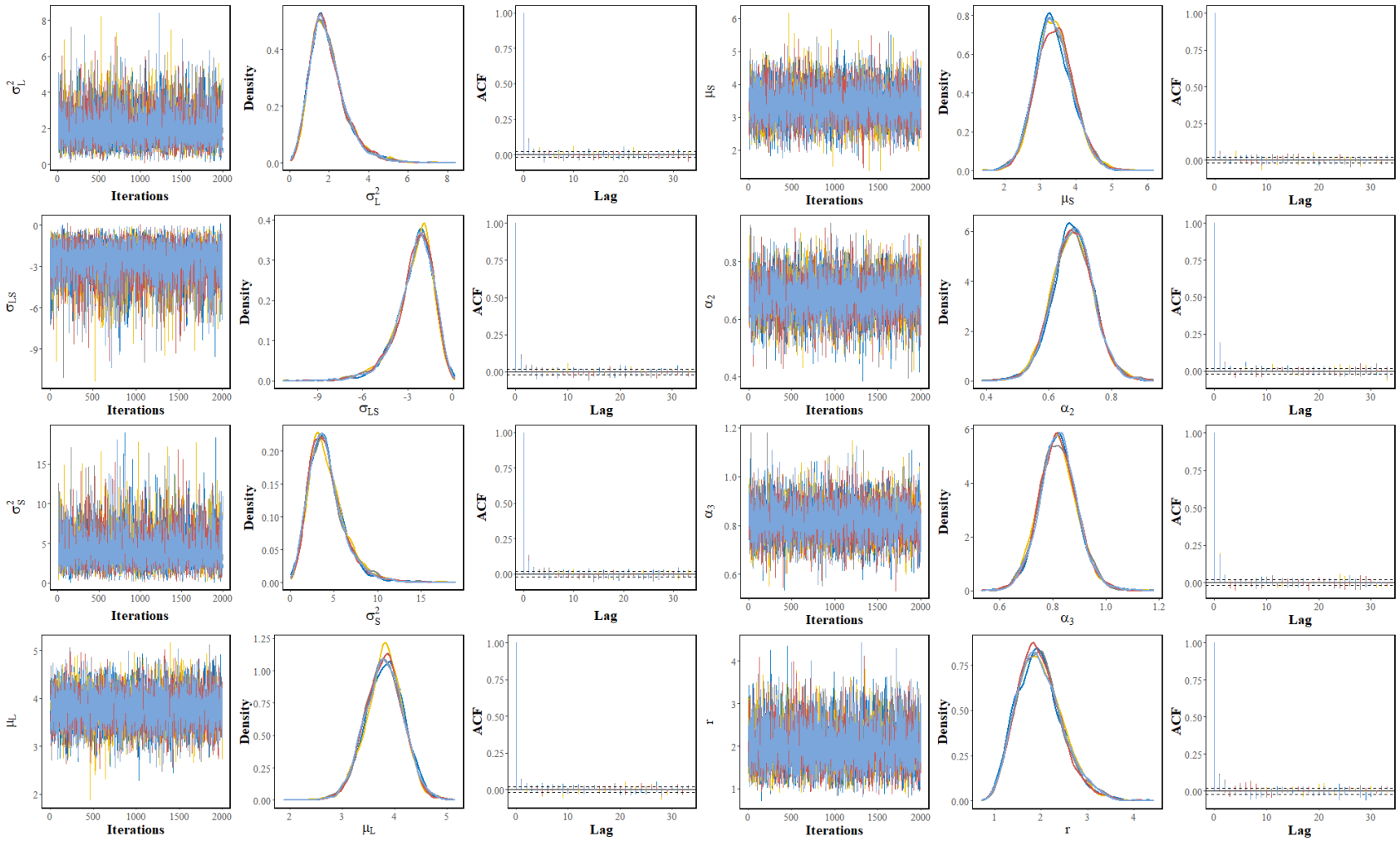

Figure D.3: Trace plots, density plots and ACF for the parameters involved on the modelling of *Total Abundance*.

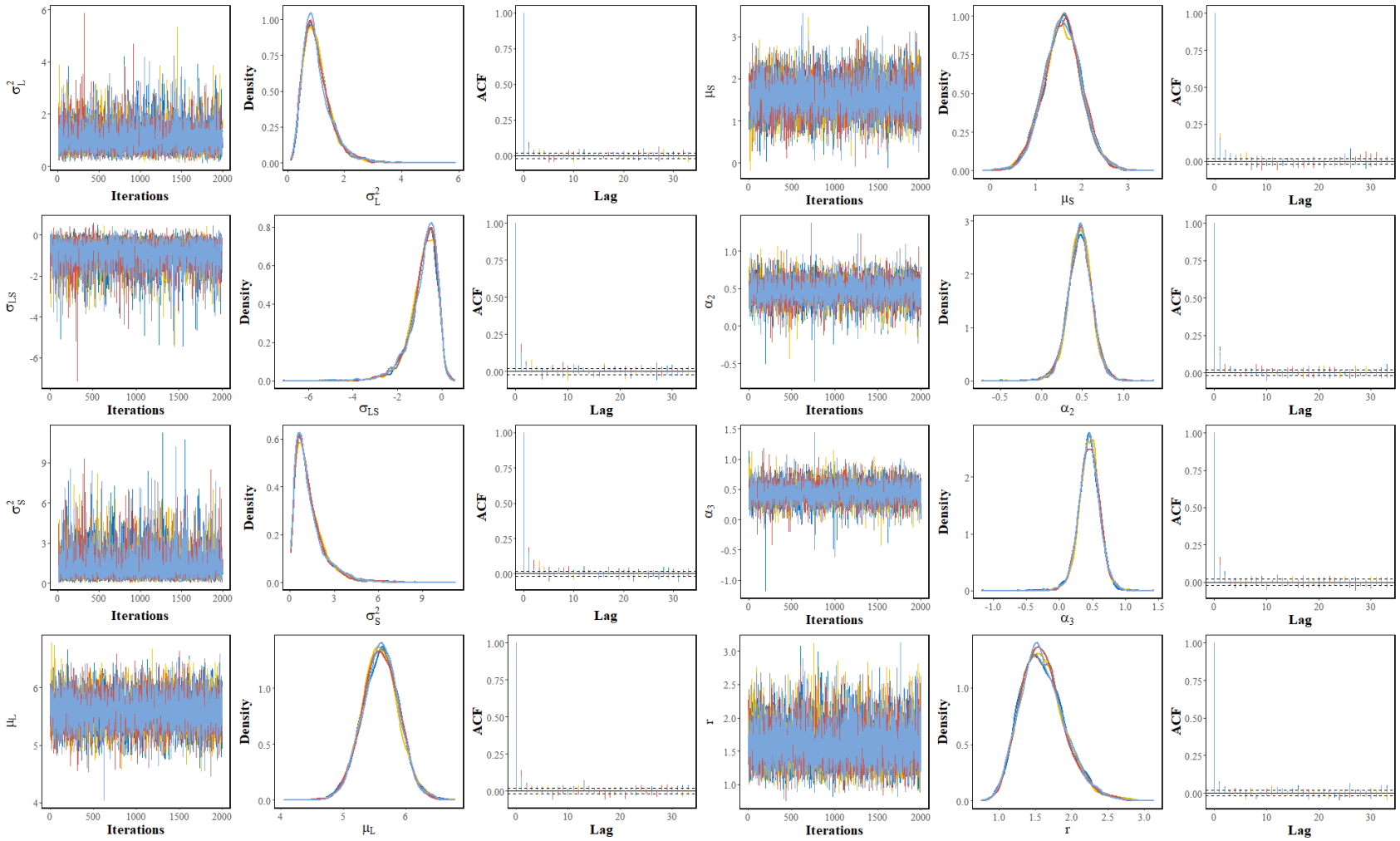

Figure D.4: Trace plots, density plots and ACF for the parameters involved on the modelling of *Total Abundance*.

D.2 *Taxonomic Richness*

Winter

Table D.5: Posterior statistics of the parameters of the models considered for the *Taxonomic Richness*' model over all sampling stations (sd - Posterior Standard Deviation, ETP - Equal Tail Probability, ESS - Effective Sample Size).

	mean	**sd**	**ETP (2.5%)**	**ETP (97.5%)**	**ESS**	$\hat{R}$
σ_L^2	0.165	0.067	0.056	0.295	10000	1.000
σ_{LS}	-0.076	0.061	-0.199	0.025	10000	1.000
σ_S^2	0.187	0.085	0.064	0.353	10000	1.000
μ_L	3.184	0.109	2.974	3.397	10000	0.999
μ_S	0.547	0.128	0.297	0.798	9438	0.999
α_2	0.369	0.151	0.080	0.668	10000	1.000
α_3	0.948	0.153	0.667	1.263	10000	0.999
r	13.384	3.506	7.332	20.286	10000	0.999

Spring

Table D.6: Posterior statistics of the parameters of the models considered for the *Taxonomic Richness*' model over all sampling stations (sd - Posterior Standard Deviation, ETP - Equal Tail Probability, ESS - Effective Sample Size).

	mean	**sd**	**ETP (2.5%)**	**ETP (97.5%)**	**ESS**	$\hat{R}$
σ_L^2	0.168	0.072	0.064	0.314	9241	0.999
σ_{LS}	-0.095	0.082	-0.274	0.026	8367	1.000
σ_S^2	0.277	0.137	0.075	0.542	8099	1.001
μ_L	3.653	0.112	3.426	3.865	8333	0.999
μ_S	0.380	0.146	0.109	0.682	8176	1.000
α_2	-1.122	0.496	-2.129	-0.138	6173	1.000
α_3	0.807	0.227	0.377	1.273	9444	1.000
r	5.685	1.243	3.430	8.153	9557	1.000

Summer

Table D.7: Posterior statistics of the parameters of the models considered for the *Taxonomic Richness*' model over all sampling stations (sd - Posterior Standard Deviation, ETP - Equal Tail Probability, ESS - Effective Sample Size).

	mean	**sd**	**ETP (2.5%)**	**ETP (97.5%)**	**ESS**	$\hat{R}$
σ_L^2	0.733	0.323	0.219	1.362	10000	1.000
σ_{LS}	-0.836	0.402	-1.613	-0.169	10000	1.000
σ_S^2	1.345	0.591	0.390	2.524	10000	0.999
μ_L	2.568	0.210	2.164	2.981	10373	1.000
μ_S	1.578	0.283	1.031	2.145	9805	1.000
α_2	0.755	0.067	0.625	0.889	8274	1.000
α_3	0.876	0.070	0.737	1.015	8132	1.000
r	9.384	3.083	4.186	15.529	9785	0.999

Autumn

Table D.8: Posterior statistics of the parameters of the models considered for the *Taxonomic Richness*' model over all sampling stations (sd - Posterior Standard Deviation, ETP - Equal Tail Probability, ESS - Effective Sample Size).

	mean	**sd**	**ETP (2.5%)**	**ETP (97.5%)**	**ESS**	$\hat{R}$
σ_L^2	0.321	0.137	0.104	0.586	9725	1.000
σ_{LS}	-0.191	0.146	-0.495	0.037	10139	1.000
σ_S^2	0.399	0.243	0.077	0.855	9633	0.999
μ_L	3.501	0.145	3.207	3.784	9224	1.000
μ_S	0.614	0.202	0.236	1.024	8963	1.000
α_2	0.649	0.172	0.328	0.996	8579	1.001
α_3	0.200	0.181	-0.174	0.535	8875	0.999
r	8.422	2.406	4.319	13.184	9465	1.000

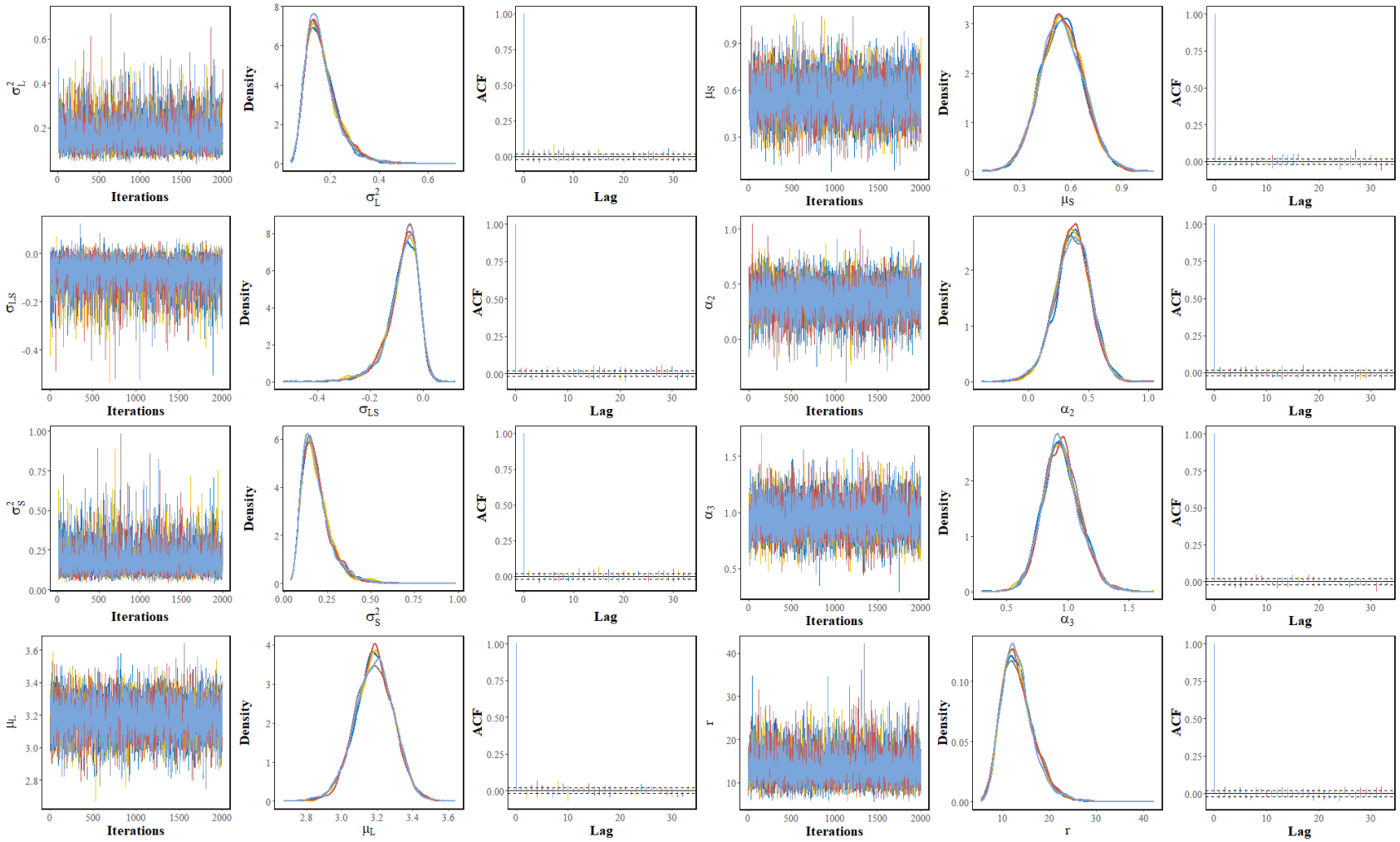

Figure D.5: Trace plots, density plots and ACF for the parameters involved on the modelling of *Taxonomic Richness*.

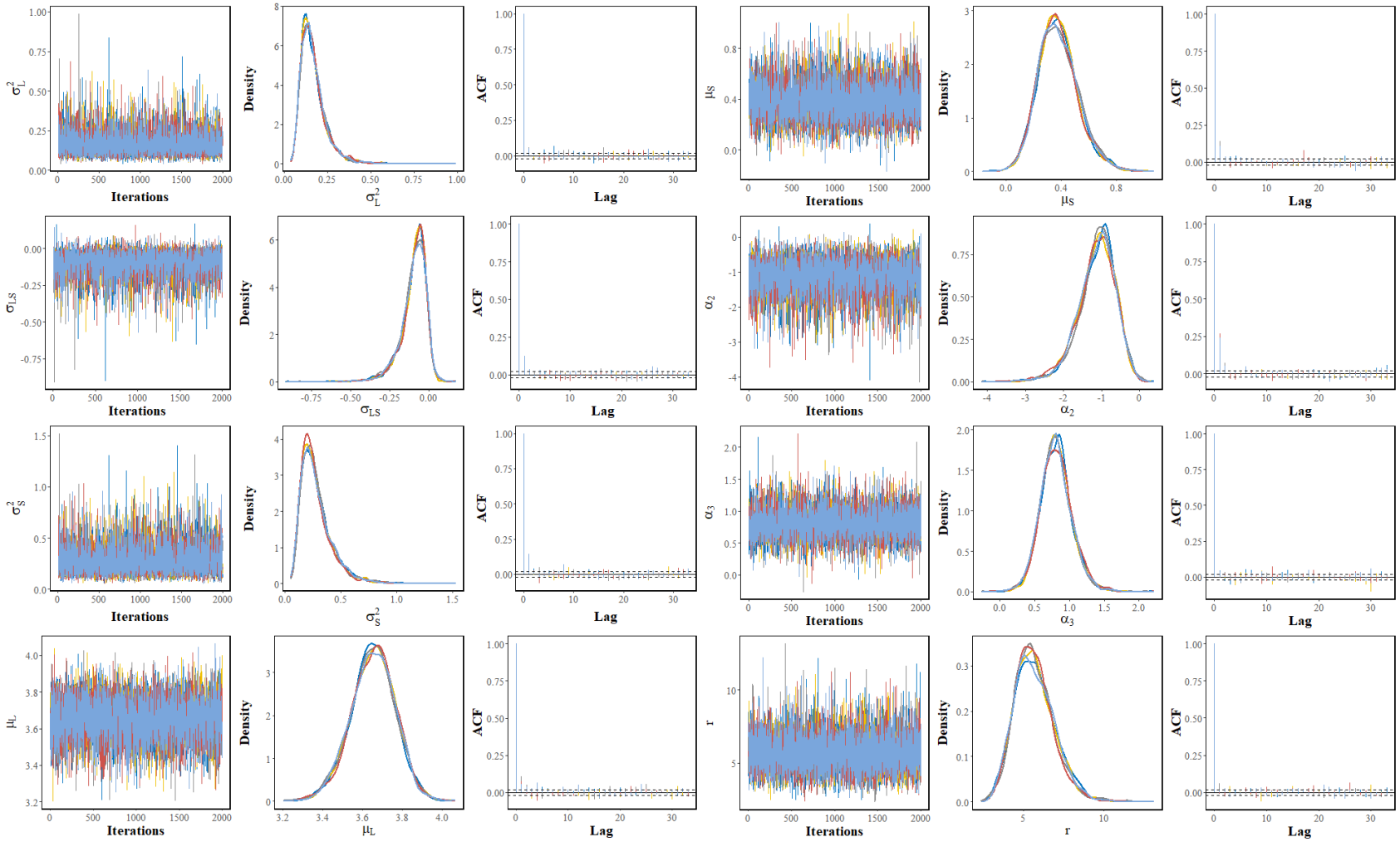

Figure D.6: Trace plots, density plots and ACF for the parameters involved on the modelling of *Taxonomic Richness*.

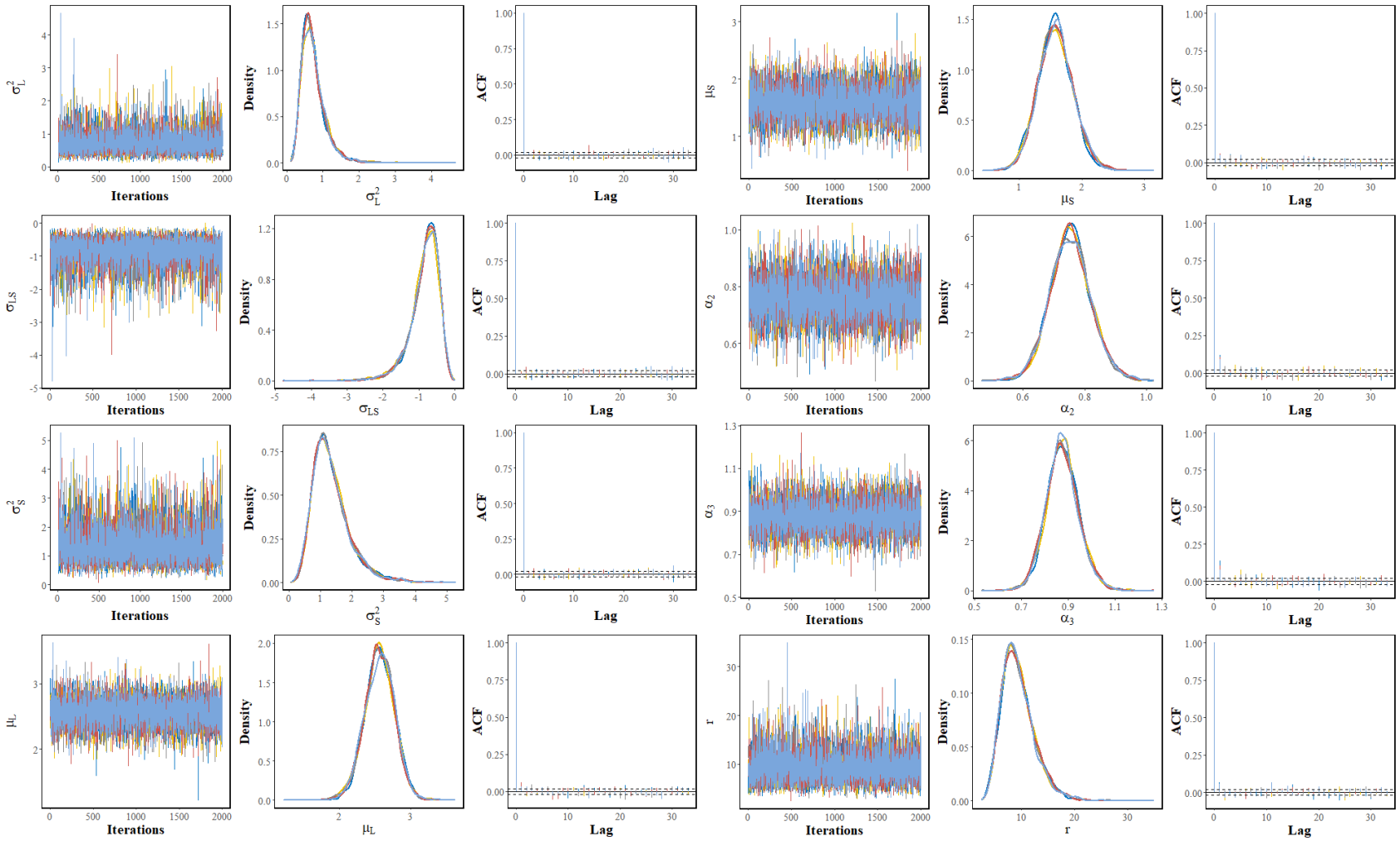

Figure D.7: Trace plots, density plots and ACF for the parameters involved on the modelling of *Taxonomic Richness*.

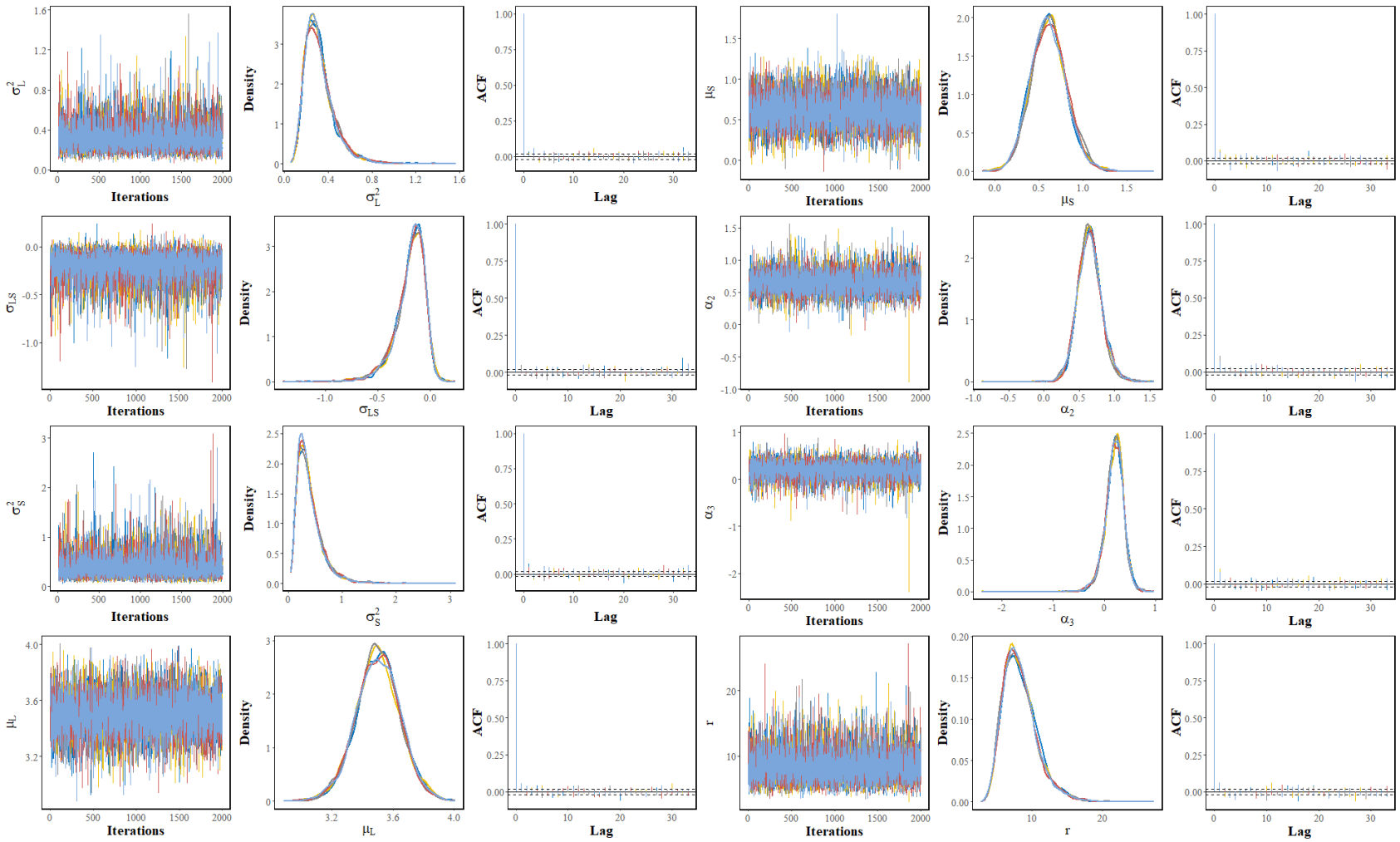

Figure D.8: Trace plots, density plots and ACF for the parameters involved on the modelling of *Taxonomic Richness*.

D.3 *Biotic Coefficient*

Winter

Table D.9: Posterior statistics of the parameters of the models considered for the *Biotic Coefficient*'s model over all sampling stations (sd - Posterior Standard Deviation, ETP - Equal Tail Probability, ESS - Effective Sample Size).

	mean	**sd**	**ETP (2.5%)**	**ETP (97.5%)**	**ESS**	$\hat{R}$
σ_L^2	0.163	0.066	0.064	0.294	9744	1.000
σ_{LS}	-0.069	0.060	-0.192	0.031	10000	0.999
σ_S^2	0.193	0.089	0.066	0.369	10000	0.999
μ_L	2.021	0.127	1.776	2.267	10000	0.999
μ_S	0.451	0.166	0.141	0.773	10000	0.999
α_2	-0.649	0.465	-1.591	0.163	10000	1.000
α_3	0.615	0.261	0.085	1.117	10000	1.000
σ_e^2	0.232	0.045	0.152	0.321	10000	0.999

Spring

Table D.10: Posterior statistics of the parameters of the models considered for the *Biotic Coefficient*'s model over all sampling stations (sd - Posterior Standard Deviation, ETP - Equal Tail Probability, ESS - Effective Sample Size).

	mean	**sd**	**ETP (2.5%)**	**ETP (97.5%)**	**ESS**	$\hat{R}$
σ_L^2	0.129	0.048	0.055	0.226	10000	0.999
σ_{LS}	-0.014	0.043	-0.101	0.070	10000	1.000
σ_S^2	0.165	0.076	0.056	0.312	9607	0.999
μ_L	2.338	0.103	2.130	2.534	10000	1.000
μ_S	0.290	0.120	0.070	0.538	10000	0.999
α_2	-0.767	0.457	-1.677	0.091	10000	1.000
α_3	1.057	0.326	0.431	1.715	10000	1.000
σ_e^2	0.190	0.037	0.124	0.262	10000	0.999

Summer

Table D.11: Posterior statistics of the parameters of the models considered for the *Biotic Coefficient*'s model over all sampling stations (sd - Posterior Standard Deviation, ETP - Equal Tail Probability, ESS - Effective Sample Size).

	mean	**sd**	**ETP (2.5%)**	**ETP (97.5%)**	**ESS**	$\hat{R}$
σ_L^2	0.263	0.143	0.069	0.540	10000	0.999
σ_{LS}	-0.204	0.176	-0.552	0.042	10000	1.000
σ_S^2	0.439	0.283	0.078	0.974	10000	1.000
μ_L	1.748	0.160	1.423	2.058	8697	0.999
μ_S	1.009	0.253	0.532	1.523	10000	0.999
α_2	0.799	0.186	0.445	1.170	10000	1.000
α_3	0.866	0.189	0.499	1.229	10000	0.999
σ_e^2	0.337	0.070	0.210	0.473	10442	1.000

Autumn

Table D.12: Posterior statistics of the parameters of the models considered for the *Biotic Coefficient*'s model over all sampling stations (sd - Posterior Standard Deviation, ETP - Equal Tail Probability, ESS - Effective Sample Size).

	mean	**sd**	**ETP (2.5%)**	**ETP (97.5%)**	**ESS**	$\hat{R}$
σ_L^2	0.190	0.087	0.062	0.361	9162	0.999
σ_{LS}	-0.079	0.091	-0.265	0.070	9117	1.000
σ_S^2	0.304	0.160	0.082	0.618	9801	1.001
μ_L	2.020	0.128	1.772	2.278	10474	1.000
μ_S	0.609	0.193	0.246	0.997	9373	1.000
α_2	0.829	0.220	0.436	1.285	10000	1.000
α_3	1.144	0.277	0.664	1.712	10000	1.000
σ_e^2	0.208	0.045	0.127	0.295	10000	0.999

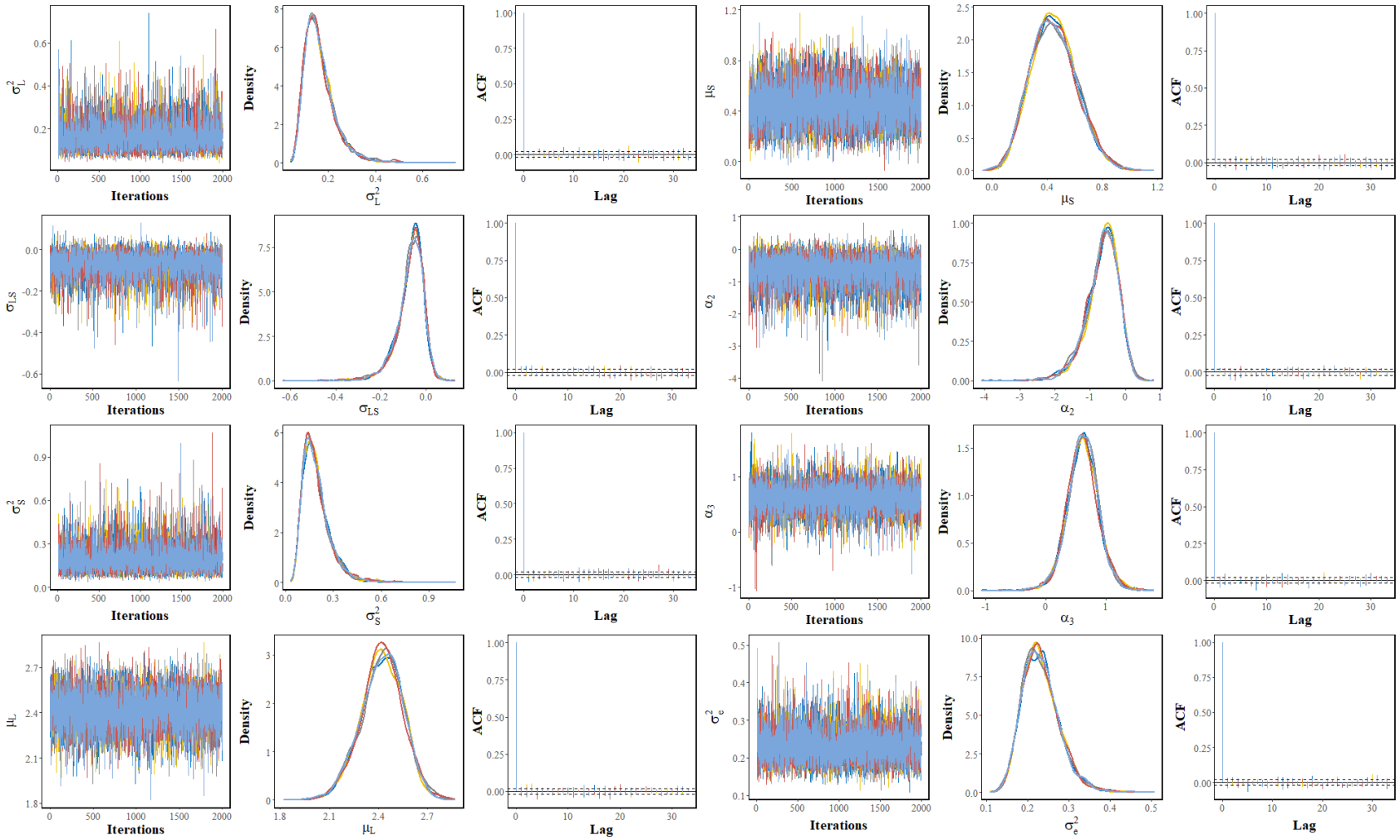

Figure D.9: Trace plots, density plots and ACF for the parameters involved on the modelling of *Biotic Coefficient*.

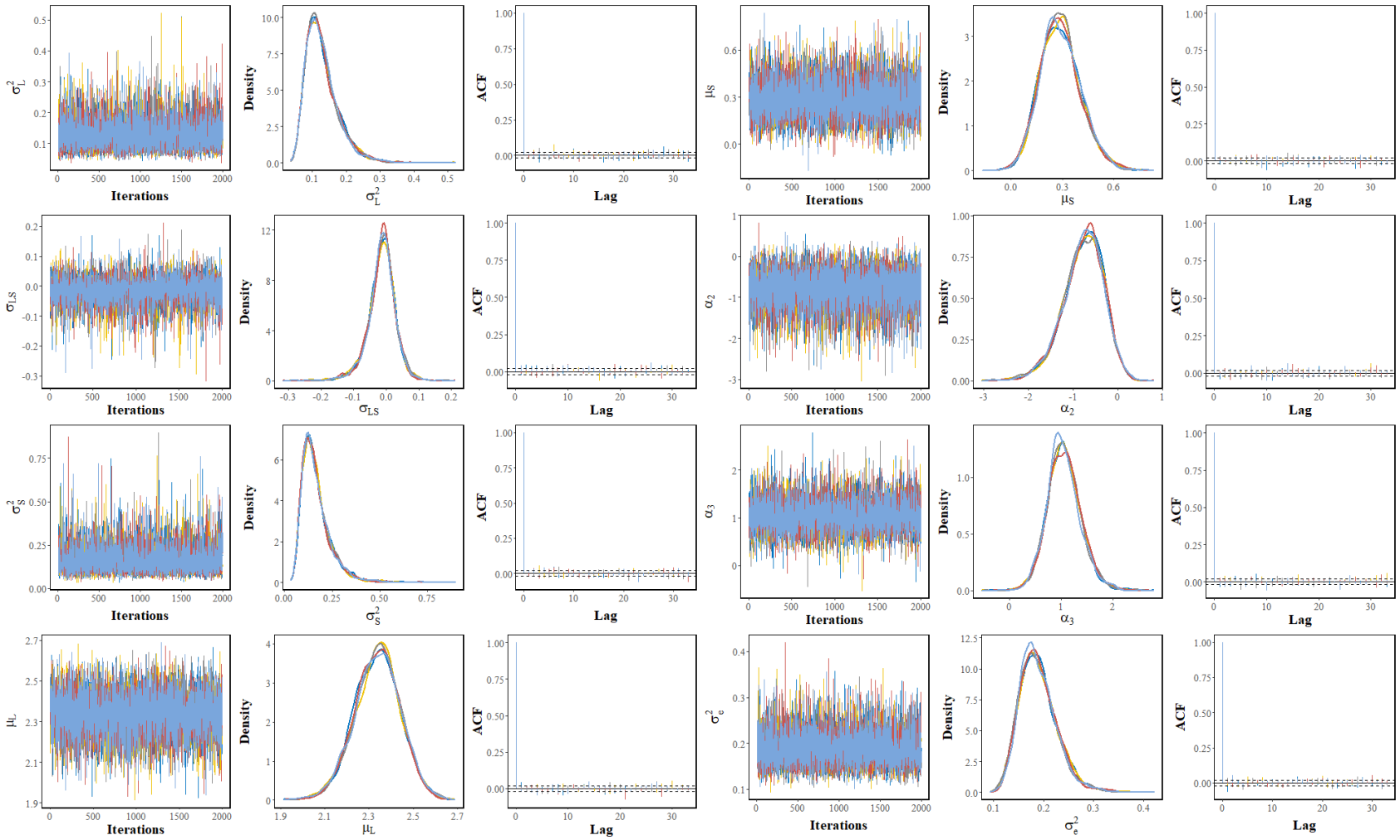

Figure D.10: Trace plots, density plots and ACF for the parameters involved on the modelling of *Biotic Coefficient*.

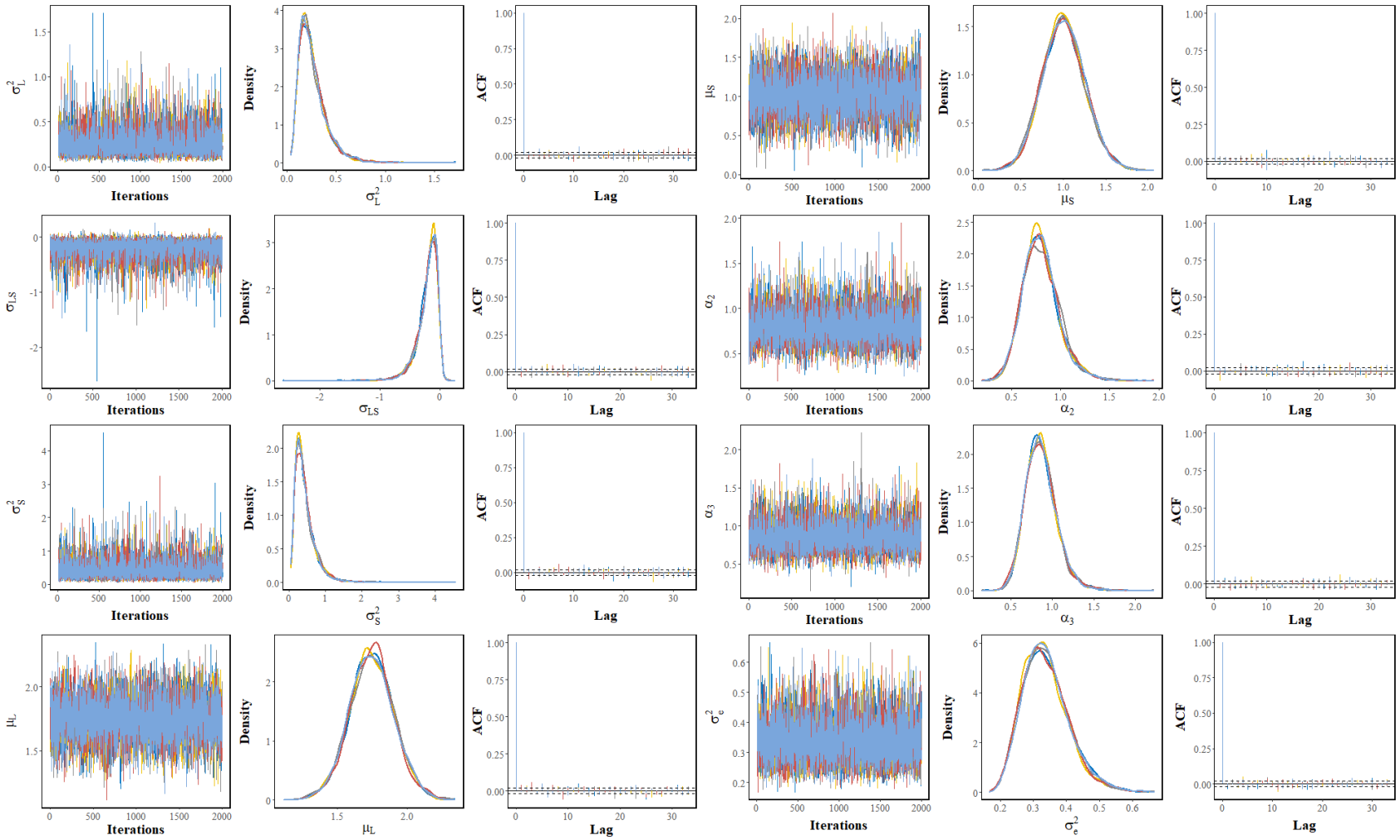

Figure D.11: Trace plots, density plots and ACF for the parameters involved on the modelling of *Biotic Coefficient*.

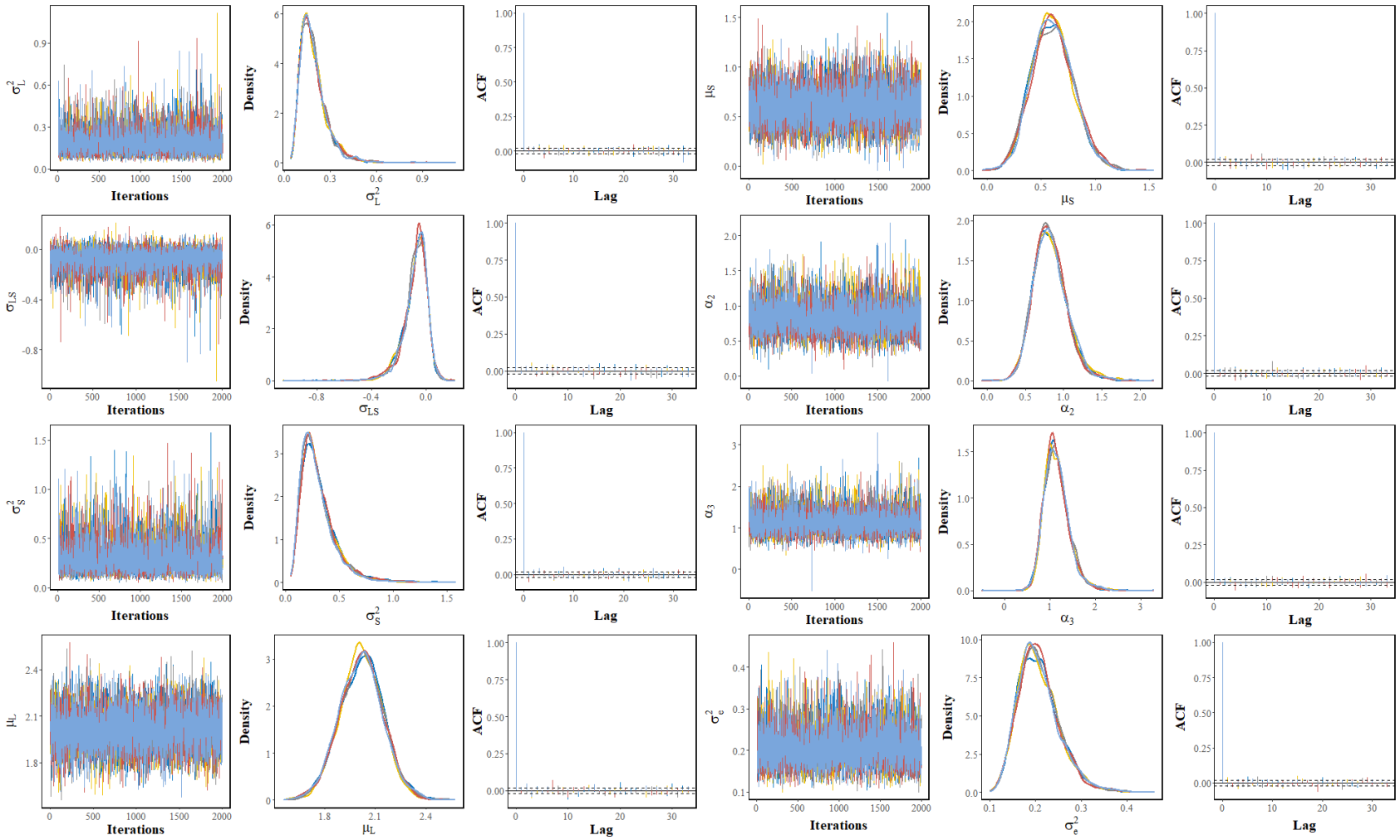

Figure D.12: Trace plots, density plots and ACF for the parameters involved on the modelling of *Biotic Coefficient*.